UNITED STATES DISTRICT COURT
DISTRICT OF NEW JERSEY

IN RE: INSULIN PRICING LITIGATION	No. 2:23-md-03080 MDL No. 3080 JUDGE BRIAN R. MARTINOTTI JUDGE RUKHSANAH L. SINGH SUPPLEMENTAL DECLARATION OF MAURA R. GROSSMAN, J.D., PH.D.

SUPPLEMENTAL DECLARATION OF MAURA R. GROSSMAN, J.D., PH.D.

I, Maura R. Grossman, declare as follows:

1. I incorporate by reference my prior declaration dated January 28, 2025, submitted in connection with the above captioned matter [ECF No. 401].

2. I continue to be retained by Plaintiffs to serve as a consultant on technology-assisted review ("TAR") related issues, particularly as relates to the development and implementation of a TAR protocol in this matter. My credentials as a TAR expert are set forth at paragraphs 4–7 of my prior declaration and in the accompanying curriculum vitae. *Id.*

3. I submit this supplemental declaration on behalf of Plaintiffs in support of the positions Plaintiffs have taken in the TAR protocol provided to Defendant Express Scripts on February 24, 2025.

4. I am generally familiar with the issues related to the TAR process and validation that have been raised in this dispute. I have personal knowledge of the matters set forth herein, and if called upon and sworn as an expert witness, I could and would competently testify thereto.

5. I understand that there continue to be material disagreements on the implementation and validation of TAR concerning three main issues: (i) training the TAR model by using *all* relevance-based coding decisions made during the course of the review process, as opposed to training it by using only selected coding

decisions made by certain attorneys, on an unknown volume of documents, based upon on undisclosed selection criteria; (ii) the use of an objective, empirically supported "stopping point" or review cut-off criterion to determine when to cease the review process and turn to validation, versus the use of a subjective, unspecified method; and (iii) the application of an unbiased, statistically sound validation protocol and recall calculation versus the application of biased, unsound methods. I will address each of these issues in turn.

I. TAR Training

6. My understanding is that Express Scripts is using Brainspace CCML, which is generally known to be a variant of the Continuous Active Learning® ("CAL®") process I co-invented with Professor Gordon V. Cormack. CAL (often referred to as "TAR 2.0") is simpler, more straightforward, and yields superior results than earlier TAR models, *e.g.*, Simple Active Learning ("SAL") or TAR 1.0.

7. Earlier, non-CAL TAR models primarily relied upon a small set of initial training documents—the "seed set"—to either rank or determine the predicted relevance of the rest of the unreviewed documents subject to TAR. This seed set normally included known responsive documents (typically identified using search terms), known non-responsive documents, and documents identified through random or non-random sampling of the TAR universe. The TAR system then ranked the document universe based on likelihood of responsiveness, or separated the TAR

universe into responsive and non-responsive subsets, based upon only those documents selected for training. This initial training was sometimes supplemented by one or more rounds of iterative training including more documents, but the training process was limited to a brief training phase of several thousand documents and did not continue throughout the TAR review process. When only a relatively small and primarily hand-selected set of documents are used for training, the choice and substance of these training documents becomes much more important because they exert a greater influence on the TAR model and its overall performance than when training is continuous throughout the TAR review.

8. The TAR 1.0 protocol described above differs significantly from CAL, which—at least as invented and promulgated by Professor Cormack and me—relies initially for training on a hand-selected seed set but thereafter uses *all* coding decisions from all documents reviewed for its training. The model is *continuously* updated and documents are continuously re-ranked based upon this ongoing training process. Documents are generally reviewed in order starting with those documents ranked most-likely-to-be responsive, but the rankings continually change as the training continues and the learning algorithm improves. This continuous training provides the TAR model with far more training examples, both in volume and diversity than the TAR 1.0 method and has been shown to be superior. *See generally* Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning*

Protocols for Technology-Assisted Review in Electronic Discovery, SIGIR '15, July 2014 (attached hereto as Exhibit A). The differences between CAL and earlier TAR protocols are outlined in detail in my paper with Gordon V. Cormack titled *Comments on 'The Implications of Rule 26(g) on the Use Technology-Assisted Review,"* Federal Courts Law Review, Volume 7, Issue 1 (2014), at 291–93 (attached hereto as Exhibit B).

9. My understanding is that Express Scripts has refused to commit to using *all*—or even the majority of—attorney coding decisions for training, as I recommend, or even to disclose when and how often they intend to retrain the TAR model. This is not *Continuous* Active Learning. As previously stated, CAL tools like the one that Express Scripts is using should re-train *continuously*, using *all* attorney relevance-based coding decisions for training rather than on only a selected, limited number of documents, following a "quality-control" process. In my opinion and based on my prior research and experience conducting hundreds of TAR reviews, the greater volume of training documents and the continuous improvement of the model are far more important to improving the TAR model—and therefore the quality of the ensuing production—than using a smaller number of "quality-controlled" documents for training. If Express Scripts is choosing to use a more selective or limited training set, then it should describe its proposed approach in detail so that it can be evaluated to ensure that the training process is not

unreasonably biased towards prioritizing narrow concepts or limited categories of documents. Beyond the selection of the initial seed set, the practice of "'cherry-picking' of training examples is a questionable practice due to its unpredictable impact on the learning algorithm." Maura R. Grossman and Gordon V. Cormack, *Comments on 'The Implications of Rule 26(g) on the Use Technology-Assisted Review,"* Federal Courts Law Review, at 299 (attached hereto as Ex. B).

10. In CAL, the process of coding the highest-ranking documents (*i.e.*, those most likely to be responsive), re-training, and then coding the next most highest-ranking documents is repeated over and over until the number of top-ranked documents containing responsive information drops precipitously, typically to 10% or less. This point is referred to as "marginal precision."

11. When implementing CAL in the intended manner, the choice of initial training documents (or the seed set) is unlikely to unreasonably bias the review process because there is sufficient training from other documents as more and more documents move through the TAR review queue. Minor errors or inconsistencies in coding have a minor impact on the TAR model as long as the overall quality of the review remains sufficiently high. However, if the majority of the training is implemented based upon a more limited set of "quality-control" documents, chosen in an undisclosed manner and an unknown (but clearly lesser) volume than *all* documents, this is not only no longer a CAL process, but it is also not a systematic,

reliable, or repeatable process. The documents selected for training and their coding become far more influential, with the potential to unreasonably bias the prioritization of the documents and the sufficiency of the production.

II. TAR Stopping Criterion

12. Express Scripts has not disclosed the criterion it will use to determine when the TAR review process may cease and validation should be undertaken. Instead, Express Scripts focuses exclusively on validation to ensure a quality TAR review. But determining when to stop the TAR review is critical to the efficiency and effectiveness of the TAR process. If the TAR review is stopped too early, a material number of responsive documents are left behind that could readily have been found. Conversely, if the TAR review is not stopped at a reasonable point, the review becomes disproportionate because an insufficient number of new and different responsive documents are being identified to justify the cost of continuing the review process.

13. The stopping point Plaintiffs have proposed—a marginal precision of 10%, meaning that the last 1,000 documents reviewed contain 10% or fewer responsive documents—is objective, reasonable, scientifically sound, and requires no extra or special effort. It has been shown, through peer-reviewed scientific study, to be a reliable predictor of when the various different aspects of relevance have all been identified and the TAR review will likely pass the validation test (in other

words, the recall estimate will be sufficiently high to suggest an adequate production). *See* Cormack and Grossman, *Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review*, SIGIR '15, Aug. 2015, at 764–65, attached as Ex. B to my initial declaration [ECF No. 401]. Continuing the review until the 10%-or-less stopping criterion is met generally achieves not only high overall recall—referring to the proportion or percent of all responsive documents in the collection that have been found—but also high recall for the different aspects or facets of relevance, including those that are infrequent in the TAR universe. *See id.*.

14. Express Scripts' contention that having an objectively reasonable stopping criteria is unnecessary as long as the review is ultimately validated is misguided. The goal of TAR—or any review process—is to identify as many responsive documents as reasonably possible, at a proportionate cost. Validation should only be undertaken when Express Scripts has an objective reason to be confident that its review is reasonably complete. 10% marginal precision provides an easy and objective standard; Express Script has proposed no such thing.

III. TAR Validation

15. The validation protocol and estimated recall calculation proposed by Plaintiffs is reasonable and consistent with validation protocols ordered in other, similar complex litigation, including *In re Broiler Chicken Antitrust Litig.*, 2018 WL 1146371 (N.D. Ill. 2018) and *In re Uber Techs, Inc., Passenger Sexual Assault Litig.*,

No. 3:23-md-03084, ECF 524. Express Scripts has characterized their proposed validation method as being consistent with that used *In re Broiler Chicken Antitrust Litig.*, but as the Special Master that ordered that validation protocol, I can state with certainty that is not the case. While Express Scripts has provided extremely limited information about its intended validation method, it is clear that Express Scripts does not intend to follow the *Broiler Chicken* validation protocol at all. The *Broiler Chicken* validation process is based on an evaluation of *three* random samples drawn from (i) documents marked responsive by reviewers, (ii) documents marked non-responsive by reviewers, and (iii) documents that were not reviewed at all because the TAR system gave them low rankings. As best I can understand it, Express Scripts intends to sample only the latter (iii)—referred to as the "null set." This validation method is called "Elusion Sampling," and it suffers from severe limitations including bias, lack of statistical soundness, and lack of reliability.

16. The validation process applied in both *In re Broiler Chicken Antitrust Litig.* and *In re Uber Techs, Inc., Passenger Sexual Assault Litig.* rely on samples drawn from the entire universe of documents subject to TAR. The documents in the validation sample set are interspersed and reviewed and coded under precisely the same conditions. The review of those documents is "blind" so that the reviewers are unaware of the previous coding or predicted responsiveness of any document in the sample. This is necessary to ensure that the validation coding is truly independent

of the prior review process and the reviewers are not influenced or biased in their responsiveness determinations. When the reviewers are given only null-set documents to examine—documents which are almost exclusively non-responsive—they are biased towards marking all the documents non-responsive. This improperly elevates the recall estimate.

17. Express Scripts has not provided a specific method for calculating the estimated recall of its production, but because its validation sample will only be drawn from the null set (*i.e.*, documents in the TAR universe that have not been reviewed or coded), it is obvious that the documents marked responsive or non-responsive by the reviewers during the course of the TAR review will be reviewed and coded under a completely different set of circumstances, which is not comparable to, the review of the null set.

18. While Express Scripts does not explain how it intends to calculate recall, which requires an estimate of the number of responsive documents found, I am presuming that Express Scripts intends to count the documents it has produced as responsive for purposes of estimating recall. This is a highly faulty assumption, since not all produced documents will in fact be responsive. Not only do reviewers make errors, but many attachments to responsive documents are not in themselves responsive. Even if Express Scripts were to exclude the latter from its calculation

of the recall estimate, it will still inevitably result in an over-inflated estimate of recall.

19. As I have indicated, recall measures the estimated proportion (or percent) of responsive documents in the TAR universe that were successfully identified as responsive. The Elusion Sampling Express Scripts proposes to do only estimates how many relevant documents were *missed*, without providing any contextual information about how many responsive documents there were to be found in the first place. Accordingly, Elusion Sampling alone does not provide useful information about the sufficiency of a production.

20. Express Scripts' contention that Plaintiffs' proposed validation method—which includes sampling both documents previously marked responsive and those previously marked non-responsive—is intended to check or "second guess" their review attorneys' coding decisions is incorrect. It is a necessary because *it is the only way to calculate an independent and statistically valid estimate of recall*. Express Scripts should not be able to rely upon purported statistical methods to support the sufficiency of its review process but then disregard the importance of the conditions under which such methods are applied. There is no question that a recall estimate based on stratified sampling—as was performed in *In re Broiler Chicken Antitrust Litig.* and *In re Uber Techs, Inc., Passenger Sexual Assault Litig*

and is proposed by Plaintiffs here—is statistically sound, while the recall estimate based on Elusion proposed by Express Scripts is not.

21. My understanding is that Express Scripts has also not committed to other aspects of the validation protocol ordered in *In re Broiler Chicken Antitrust Litig.*, such as confirming that they will promptly provide Plaintiffs with the responsive documents identified through the validation process. Plaintiffs are obviously entitled under the Federal Rules of Civil Procedure to see all non-privileged responsive documents. But while Express Scripts balks at establishing any pre-set, *quantitative* recall target, they also refuse to provide Plaintiffs with the very *qualitative* information they would need to ensure that unique or significant documents have not eluded production. Frankly, it surprises me that they cite approvingly my *eDiscovery Medicine Show* paper with Professor Cormack (*see* 18:1 Ohio State Tech. L. J. 1 (2021)) (attached hereto as Exhibit C)—at footnote 4 of their Feb. 5, 2025 letter to the Court [ECF No. 410]. While in that paper, I do argue that recall alone is not the end-all and be-all indicator of an adequate production and "should be considered holistically"—meaning that *the importance of the documents missed should be considered alongside their volume*—in that paper, I also strongly eschew the use of non-blind validation and the misapplication of statistical methods, both of which Express Scripts want to do here. ("[I]f the samples are assessed by Clever Hanses who know *or can intuit* how the documents were previously coded .

. . over- or underestimation [of recall] can result"; "[P]roducing parties should show—and the courts should demand that they show—the reasonableness of their eDiscovery search and review processes, as well a the resulting production, by hewing closely to tools, methods, and procedures that have been scientifically vetted and shown to be valid and reliable. Anything else belongs in a medicine show." *Id.* at 6, 8 (emphasis added).

22. Plaintiffs have advised me that they have made a compromise offer to lower the "target recall" from 80% to 70-80% or above. That is a fair and appropriate target for recall when it is calculated on an independent, blind, stratified sample of all documents in the TAR universe and included or excluded from production.

I declare under penalty of perjury that the foregoing is true and correct to the best of my knowledge.

Executed on February 28, 2025, in Waterloo, Ontario, Canada.

MAURA R. GROSSMAN, J.D., PH.D.

EXHIBIT A

Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman*
Wachtell, Lipton, Rosen & Katz
mrgrossman@wlrk.com

ABSTRACT

Using a novel evaluation toolkit that simulates a human reviewer in the loop, we compare the effectiveness of three machine-learning protocols for technology-assisted review as used in document review for discovery in legal proceedings. Our comparison addresses a central question in the deployment of technology-assisted review: Should training documents be selected at random, or should they be selected using one or more non-random methods, such as keyword search or active learning? On eight review tasks – four derived from the TREC 2009 Legal Track and four derived from actual legal matters – recall was measured as a function of human review effort. The results show that entirely non-random training methods, in which the initial training documents are selected using a simple keyword search, and subsequent training documents are selected by active learning, require substantially and significantly less human review effort ($P < 0.01$) to achieve any given level of recall, than passive learning, in which the machine-learning algorithm plays no role in the selection of training documents. Among passive-learning methods, significantly less human review effort ($P < 0.01$) is required when keywords are used instead of random sampling to select the initial training documents. Among active-learning methods, continuous active learning with relevance feedback yields generally superior results to simple active learning with uncertainty sampling, while avoiding the vexing issue of "stabilization" – determining when training is adequate, and therefore may stop.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Search process, relevance feedback.

Keywords: Technology-assisted review; predictive coding; electronic discovery; e-discovery.

*The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2609601.

1. INTRODUCTION

The objective of technology-assisted review ("TAR") in electronic discovery ("e-discovery")[1] is to find as nearly all of the relevant documents in a collection as possible, with reasonable effort. While this study does not presume to interpret the common-law notion of what is reasonable, it serves to quantify the tradeoff between how nearly all of the relevant documents can be found (as measured by recall), and the human effort needed to find them (as measured by the number of documents that must be manually reviewed, which translates into time and cost).

In a typical review task, a "requesting party" prescribes relevance[2] by way of a "request for production," while the "responding party," its adversary, is required to conduct a review and produce the responsive, non-privileged documents identified as a result of a reasonable search. A study by Grossman and Cormack [8] shows that two TAR methods can be both more effective and more efficient than traditional e-discovery practice, which typically consists of keyword or Boolean search, followed by manual review of the search results. One of these methods, due to Cormack and Mojdeh [7], employs machine learning in a protocol we refer to as Continuous Active Learning ("CAL"). The other method, due to H5 [13], does not employ machine learning, and therefore is not considered in this study.

Often relying on Grossman and Cormack for support, many legal service providers have advanced TAR tools and methods that employ machine learning, but not the CAL protocol. These tools and methods, often referred to in the legal marketplace as "predictive coding," follow one of two protocols which we denote Simple Active Learning ("SAL") and Simple Passive Learning ("SPL"). Some tools that employ SAL have achieved superior results at TREC, but have never, in a controlled study, been compared to CAL. Tools that use SPL, while widely deployed, have not achieved superior results at TREC, and have not, in a controlled study, been compared to traditional methods, to SAL, or to CAL.

[1]See Grossman and Cormack [10] for a glossary of terms pertaining to TAR. See Oard and Webber [16] for an overview of information retrieval for e-discovery.
[2]The IR term "relevant" generally describes a document sought by an information-retrieval effort, while the legal term "responsive" describes a document that satisfies the criteria set forth in a request for production. In this study, the terms are used interchangeably; however, in the context of litigation, relevance may take on a broader meaning than responsiveness.

This study compares CAL, SAL, and SPL, and makes available a TAR evaluation toolkit[3] to facilitate further comparisons. The results show SPL to be the least effective TAR method, calling into question not only its utility, but also commonly held beliefs about TAR. The results also show that SAL, while substantially more effective than SPL, is generally less effective than CAL, and as effective as CAL only in a best-case scenario that is unlikely to be achieved in practice.

2. THE TAR PROCESS

The TAR process, in the abstract, proceeds as follows. Given a document collection and a request for production, a human operator uses one or more tools to identify documents to be shown to one or more human reviewers, who may or may not be the same individual as the operator. The reviewers examine these documents and label ("code") them each as responsive or not. More documents are identified using the tools, reviewed and coded by reviewers, and the process continues until "enough" of the responsive documents have been reviewed and coded. How many constitute "enough" is a legal question, which is informed by how much additional effort would likely be required to find more responsive documents, and how important those documents would likely be in resolving the legal dispute (*i.e.*, "proportionality considerations"). For our purposes, we consider the process to continue indefinitely, and track the number of responsive documents found (*i.e.*, recall) as a function of effort (*i.e.*, the number of documents reviewed and coded by reviewers). Using this information, the reader can determine retrospectively, for any definition of "enough," how much effort would have sufficed to find enough documents.

For this study, the operator is assumed to follow a strict protocol. All choices, including what tools are used, and when and how they are used, are prescribed by the protocol. In addition to satisfying the requirements for a controlled comparison, the use of a strict protocol may be appealing in the e-discovery context because the requesting party may distrust, and therefore wish to prohibit discretionary choices made by the operator on behalf of the responding party. The reviewers are assumed to code the documents they review in good faith, to the best of their abilities. In light of Grossman and Cormack [8, 9], and others [4, 18, 25, 26], it is unrealistic to assume the reviewers to be infallible – they will necessarily, but inadvertently, code some responsive documents as non-responsive, and vice versa.

The CAL protocol involves two interactive tools: a keyword search system and a learning algorithm. At the outset of the TAR process, the operator typically uses a keyword search to identify an initial set of documents to be reviewed and coded. These coded documents (often referred to as the "seed set") are used to train a learning algorithm, which scores each document in the collection by the likelihood that it is responsive. The top-scoring documents that have not yet been coded are then reviewed and coded by reviewers. The set of all documents coded thus far (the "training set") is used to train the learning algorithm, and the process of selecting the highest-scoring documents, reviewing and coding them, and adding them to the training set continues until "enough" of the responsive documents have been found.

[3]Available at `http://cormack.uwaterloo.ca/cormack/tar-toolkit`.

The SAL protocol, like CAL, begins with the creation of a seed set that is used to train a learning algorithm. The seed set may be selected using keywords, random section, or both, but, unlike CAL, the subsequent training documents to be reviewed and coded are selected using uncertainty sampling [15], a method that selects the documents about which the learning algorithm is least certain. These documents are added to the training set, and the process continues until the benefit of adding more training documents to the training set would be outweighed by the cost of reviewing and coding them (a point often referred to as "stabilization"). At this point, the learning algorithm is used for the last time to create either a set or a ranked list of likely relevant documents (the "review set"), which is subsequently reviewed and coded by reviewers.

The SPL protocol, unlike CAL or SAL, generally relies on the operator or random selection, and not the learning algorithm, to identify the training set. The process is typically iterative. Once a candidate training set is identified, the learning algorithm is then trained on these documents and used to create a candidate review set. If the review set is "inadequate," the operator creates a new candidate training set, generally by adding new documents that are found by the operator, or through random selection. The process continues until the review set is deemed "adequate," and is subsequently reviewed and coded by reviewers.

The TAR process addresses a novel problem in information retrieval, which we denote here as the "TAR Problem." The TAR Problem differs from well-studied problems in machine learning for text categorization [21] in that the TAR process typically begins with no knowledge of the dataset and continues until most of the relevant documents have been identified and reviewed. A classifier is used only incidentally for the purpose of identifying documents for review. Gain is the number of *relevant* documents presented to the human during training and review, while cost is the total number of *relevant and non-relevant* documents presented to the human during training and review.

3. SIMULATING REVIEW

To simulate the application of a TAR protocol to a review task, we require a realistic document collection and request for production, a keyword query ("seed query") to be used (if required by the protocol), and a simulated reviewer. To evaluate the result, we require a "gold standard" indicating the true responsiveness of all, or a statistical sample, of the documents in the collection.

Four review tasks, denoted Matters 201, 202, 203, and 207, were derived from Topics 201, 202, 203, and 207 of the TREC 2009 Legal Track Interactive Task – the same Topics that were used to evaluate Cormack and Mojdeh's CAL efforts at TREC [7, 12]. Four other review tasks, denoted Matters A, B, C, and D, were derived from actual reviews conducted in the course of legal proceedings. Statistics for the collections are provided in Table 1, and the requests for production are shown in Table 2.

The seed queries for the tasks derived from TREC, shown in Table 3, were composed by Open Text in the course of its participation in the TREC 2010 Legal Track Learning Task (which used the same topics as the TREC 2009 Legal Track Interactive Task), using a strategy that "attempted to quickly create [a] Boolean query for each topic" [24, page 5]. The seed queries for the tasks derived from legal proceedings,

Matter	Collection Size	# Rel. Docs.	Prevalence (%)
201	723,537	2,454	0.34
202	723,537	9,514	1.31
203	723,537	1,826	0.25
207	723,537	8,850	1.22
A	1,118,116	4,001	0.36
B	409,277	6,236	1.52
C	293,549	1,170	0.48
D	405,796	15,926	3.92

Table 1: Collection statistics.

Matter	Request for Production
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as *prepay transactions*.
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
207	All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.
A	[Regulatory request]
B	[Regulatory request]
C	[Third-party subpoena]
D	[Regulatory request]

Table 2: Requests for production.

described in Table 3, were composed by or negotiated with the requesting party prior to the review process. The recall and precision of each of the seed queries (as measured with respect to the gold standard, discussed below) are shown in Table 3.

To simulate a reviewer, we use a "training standard" that consists of a relevance assessment for each document in the collection. If, during the course of simulating a particular review protocol, the reviewer is called upon to code a document, the assessment from the training standard – responsive or not responsive – is used for this purpose. The training standard does not represent ground truth; instead, it represents the coding decision that a fallible reviewer might render when presented with the document for review. For all of the simulated tasks, all of the positive training-standard assessments, and some of the negative assessments, were rendered by a reviewer during the course of a prior review. For the TREC-derived tasks, we used Cormack and

Matter	Seed Query	Recall	Prec.
201	`"pre-pay" OR "swap"`	0.436	0.038
202	`"FAS" OR "transaction" OR "swap" OR "trust" OR "Transferor" OR "Transferee"`	0.741	0.090
203	`"forecast" OR "earnings" OR "profit" OR "quarter" OR "balance sheet"`	0.872	0.034
207	`"football" OR "eric bass"`	0.492	0.167
A	`[7-term Boolean query]`	0.545	0.045
B	`[98-term Boolean query]`	0.991	0.019
C	`[46-term Boolean query]`	0.259	0.026
D	`[9-term Boolean query]`	0.945	0.325

Table 3: Keyword seed queries and their associated recall and precision.

Matter	Training Standard	
	Recall	Precision
201	0.843	0.911
202	0.844	0.903
203	0.860	0.610
207	0.896	0.967
A	1.000	0.307
B	0.942	0.974
C	1.000	0.429
D	0.961	1.000

Table 4: Recall and precision for the training standard used to simulate human review.

Mojdeh's TREC submissions;[4] for the legal-matter-derived tasks, we used the coding rendered by the first-pass reviewer in the course of the review. Documents that were never seen by the first-pass reviewer (because they were never identified as potentially responsive) were deemed to be coded as non-responsive. Overall, as measured with respect to the gold standard, the recall and precision of the training standard (shown in Table 4) indicate that the simulated reviewer achieves a high-quality – but far from perfect – result, by human review standards.

In contrast to the training standard, the gold standard represents ground truth. For the TREC-derived tasks, the gold standard consists of a stratified random sample, assessed by TREC using a two-stage adjudication process [12]. For the legal-matter-derived tasks, the gold standard consists of the documents produced to the requesting party, after a second-pass review and quality-assurance efforts. Each document in the gold standard is associated with an inclusion probability – the prior probability that it would have been included in the gold standard. Following TREC practice, the recall of any simulated review is estimated using the Horvitz-Thompson estimator [14], which weights each gold-standard document by the reciprocal of its inclusion probability.

Evaluation results are presented in two ways: as gain curves and as 75% recall-effort ("75% RE") values. A gain curve plots recall as a function of the number of documents reviewed. For any level of effort (as measured by the number

[4]Available at `http://trec.nist.gov/results.html`, subject to a usage agreement.

of documents reviewed), one can determine at a glance the recall that would be achieved, using a particular protocol, for that level of effort (see Figures 1 and 2 below). Conversely, for any recall level it is possible to determine what level of effort would be required to achieve that recall level. For the purpose of quantitative comparison, we tabulate 75% RE for all protocols (see Tables 5 and 6 below).

4. TAR PROTOCOLS

In this study, we used the same feature engineering and learning algorithm for every protocol, without any collection- or task-specific tuning. Following Cormack and Mojdeh [7], the first 30,000 bytes of the ASCII text representation of each document (including a text representation of the sender, recipient, cc or bcc recipients, subject, and date and time sent) was shingled as overlapping 4-byte segments. The number of distinct possible segments was reduced, by hashing, from $2^{32} = 4,294,967,296$ to $1,000,081$ (an arbitrarily chosen prime number near one million). Each feature consisted of a binary value: "1" if the feature was present in the first 30,000 bytes of the document; "0" if it was absent. For the learning algorithm, we used the Sofia-ML implementation of Pegasos SVM,[5] with the following parameters: "`--iterations 2000000 --dimensionality 1100000`."

For all protocols, we used a batch size of 1,000 documents. That is, the initial training set (the seed set) was 1,000 documents, and each iteration, whether CAL, SAL, or SPL, involved reviewing 1,000 documents and, if indicated by the protocol, adding them to the training set. Our primary experiments evaluated the specific formulations of CAL, SAL, and SPL described in Section 2; secondary experiments explored the effect of using keyword-selected versus randomly selected documents for the seed and training sets.

Our primary CAL implementation used, as the initial training set, 1,000 documents, randomly selected from the results of a search using the seed query. In each iteration, the training-set documents were coded according to the training standard, then used to train Sofia-ML, and hence to score the remaining documents in the collection. The 1,000 top-scoring documents were added to the training set, and the process was repeated 100 times.

Our primary SAL implementation used exactly the same 1,000-document keyword-selected seed set as CAL. Like CAL, in each iteration, the training-set documents were coded according to the training standard, then used to train Sofia-ML, and hence to score the remaining documents in the collection. These documents, ranked by score, constitute a candidate review set. Rather than implementing the decision as to whether stabilization had occurred, we recorded the candidate review set for future retrospective evaluation, and continued the training process. Unlike CAL, the 1,000 documents with the *least magnitude* scores were coded and added to the training set, and the process was repeated 100 times. In the end, the simulation yielded 100 different candidate review sets, corresponding to stabilization having occurred with a training-set size of 1,000, 2,000, ..., 100,000 documents. Each training-set size, when evaluated, yields a different gain curve, and a different 75% RE. Due to space considerations, we show gain curves only for the representative training-set sizes of 2,000, 5,000, and 8,000 documents. We report 75% RE for these three training-set sizes, as well as for the *ideal* training-set size, which in reality would not be known, since it requires the benefit of hindsight. The ideal training-set size is derived using the gold standard; 75% RE is calculated for every training-set size, and the lowest value is chosen.

Our primary SPL implementation used random selection throughout, as advocated by some SPL proponents. The initial training set (which we denote the "seed set," notwithstanding the fact that many SPL proponents use the same term to refer to the final training set) consisted of 1,000 randomly selected documents, and each iteration added 1,000 more randomly selected documents. As for SAL, we recorded the candidate review set after each iteration, and report gain curves for the representative training-set sizes of 2,000, 5,000, and 8,000 documents, as well as 75% RE for these training-set sizes, and for the *ideal* training-set size, as defined above.

Variants of these protocols, for which we report 75% RE, include using randomly selected documents as a seed set for CAL and SAL, using a keyword-selected seed set for SPL, and using an entirely keyword-selected training set for SPL.

5. PRIMARY RESULTS

As illustrated in Figure 1, the CAL protocol achieves higher recall than SPL, for less effort, for all of the representative training-set sizes. All eight graphs show the same basic result: After the first 1,000 documents (*i.e.*, the seed set), the CAL curve shows a high slope that is sustained until the majority of relevant documents have been identified. At about 70% recall, the slope begins to fall off noticeably, and effectively plateaus between 80% and 100% recall. The SPL curve exhibits a low slope for the training phase, followed by a high slope, falloff, and then a plateau for the review phase. In general, the slope immediately following training is comparable to that of CAL, but the falloff and plateau occur at substantially lower recall levels. While the initial slope of the curve for the SPL review phase is similar for all training-set sizes, the falloff and plateau occur at higher recall levels for larger training sets. This advantage of larger training sets is offset by the greater effort required to review the training set: In general, the curves for different training sets cross, indicating that a larger training set is advantageous when high recall is desired.

75% recall effort, shown in Table 5, illustrates the superiority of CAL over SPL, even when SPL is afforded the benefit of hindsight to choose the ideal training-set size. A simple sign test shows with statistical significance ($P < 0.01$) that CAL is superior to SPL according to 75% RE, and also according to recall effort for any other level of recall.

Figure 2 shows that the CAL protocol generally achieves higher recall than SAL. However, the SAL gain curves, unlike the SPL gain curves, often touch the CAL curves at one specific inflection point. The strong inflection of the SAL curve at this point is explained by the nature of uncertainty sampling: Once stabilization occurs, the review set will include few documents with intermediate scores, because they will have previously been selected for training. Instead, the review set will include primarily high-scoring and low-scoring documents. The high-scoring documents account for the high slope before the inflection point; the low-scoring documents account for the low slope after the inflection point; the absence of documents with intermediate scores accounts for the sharp transition. The net effect

[5]Available at `http://code.google.com/p/sofia-ml`.

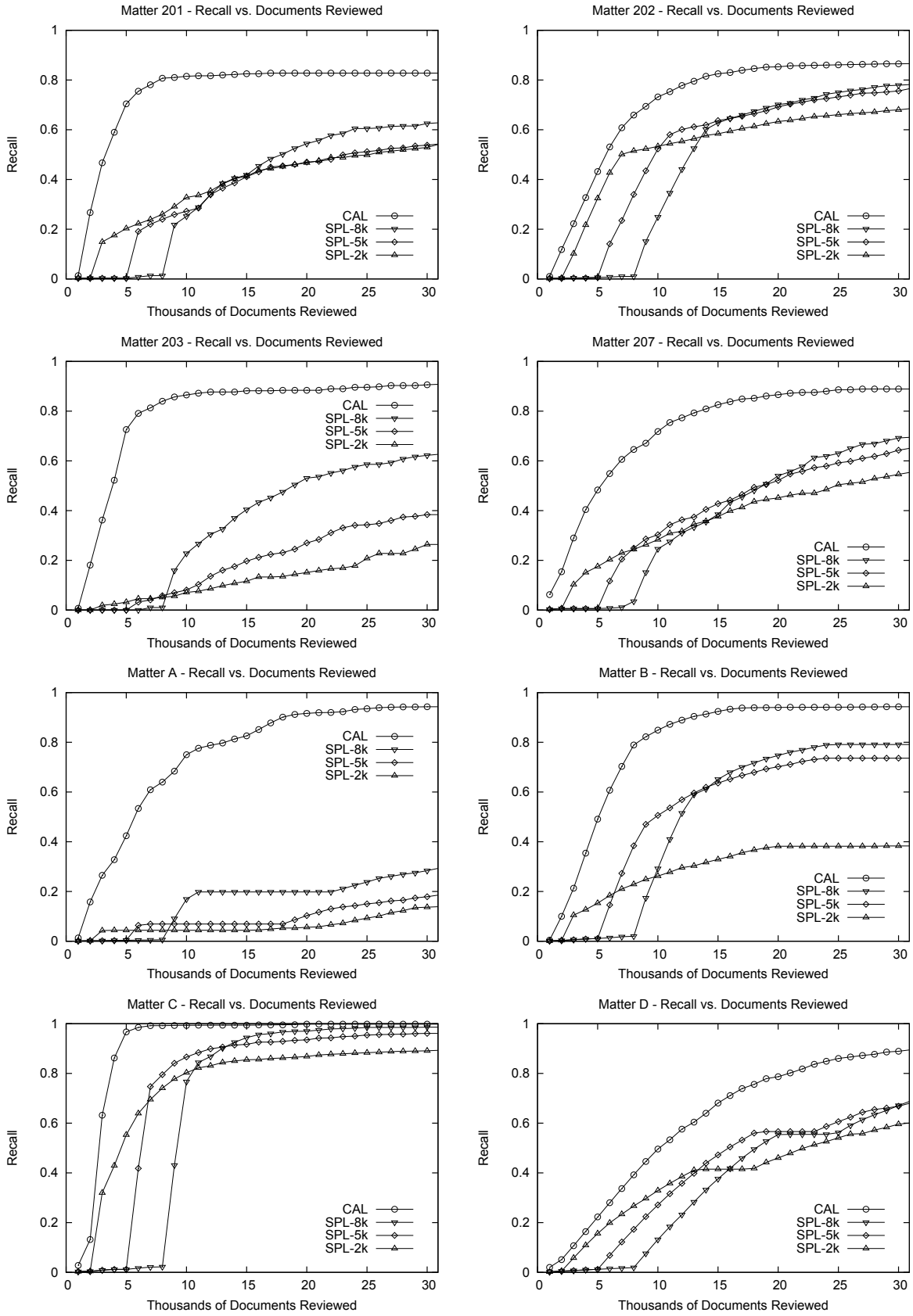

Figure 1: Continuous Active Learning versus Simple Passive Learning using three different training-set sizes of randomly selected documents.

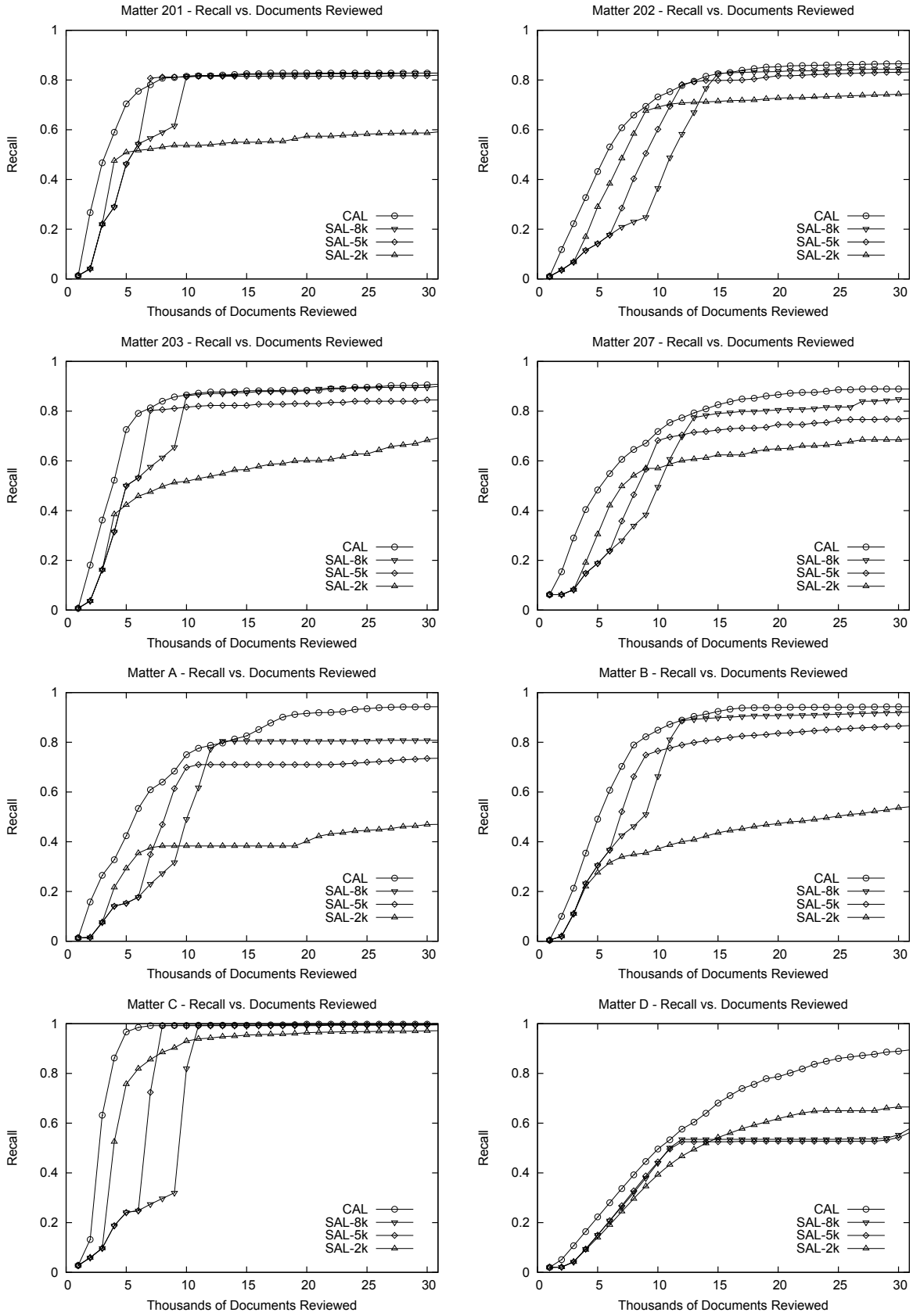

Figure 2: Continuous Active Learning versus Simple Active Learning using three different training-set sizes of uncertainty-sampled documents.

Matter	CAL	SAL				SPL			
		Training Set Size				Training Set Size			
		2K	*5K*	*8K*	*Ideal*	*2K*	*5K*	*8K*	*Ideal*
201	**6**	237	7	10	7	284	331	164	56
202	**11**	34	12	14	12	47	29	26	26
203	**6**	43	7	10	**6**	521	331	154	99
207	**11**	55	23	13	13	103	50	36	35
A	**11**	210	42	12	12	502	326	204	85
B	**8**	119	10	11	10	142	41	21	20
C	**4**	5	8	10	5	9	8	10	7
D	**18**	60	54	53	**18**	55	38	37	37

Table 5: 75% Recall Effort for Primary Results (measured in terms of thousands of documents reviewed). Bold numbers reflect the least possible effort to achieve the target recall of 75%.

Matter	CAL	CAL-seedran	SAL	SAL-seedran	SPL	SPL-seedkey	SPL-allkey
			Training Set Size		Training Set Size		
			Ideal	*Ideal*	*Ideal*	*Ideal*	*Ideal*
201	**6**	**6**	7	8	56	36	43
202	**11**	12	12	12	26	23	20
203	**6**	637	**6**	614	99	26	16
207	**11**	12	13	13	35	26	16
A	**11**	15	12	15	85	79	66
B	**8**	10	10	10	20	19	39
C	**4**	**4**	5	**4**	7	6	9
D	**18**	19	**18**	19	37	28	34

Table 6: 75% Recall Effort for Primary and Supplemental Results (measured in terms of thousands of documents reviewed). Bold numbers reflect the least possible effort to achieve the target recall of 75%.

is that SAL achieves effort as low as CAL only for a specific recall value, which is easy to see in hindsight, but difficult to predict at the time of stabilization.

Table 5 illustrates the sensitivity of the SAL and SPL results to training-set size, and hence the difficulty of choosing the precise training-set size to achieve 75% recall with minimal effort.

6. SUPPLEMENTAL RESULTS

To assess the role of keyword versus random selection at various stages of the training process, we evaluated the following variants of the primary protocols: (i) CAL-seedran, in which the seed set was selected at random from the entire collection; (ii) SAL-seedran, in which the seed set was selected at random from the entire collection; (iii) SPL-seedkey, in which the initial 1,000 training documents were the same keyword-selected seed set used for CAL and SAL in the primary protocols; and (iv) SPL-allkey, in which all training examples were selected at random from the results of the keyword seed query. 75% recall effort (with ideal training-set sizes, where applicable) for these variants, as well as the primary protocols, is shown in Table 6.

A comparison of the results for CAL and CAL-seedran shows that a random seed set generally yields the same or slightly inferior results to a keyword-selected seed set. In one case – Matter 203 – the random seed set fails spectacularly. The collection for this task has very low prevalence (0.25%), and the seed set of 1,000 random documents contained only two responsive documents, which were insufficient to "kick-start" the active-learning process. A comparison of the results for SAL and SAL-seedran shows the same general effect, including the degraded performance caused by random seeding for Matter 203.

A comparison of the results for SPL and SPL-seedkey shows that, as for CAL and SAL, the use of keyword selection for the *initial* training set generally yields superior results to random selection. A comparison of the results for SPL and SPL-allkey shows that, with two exceptions, keyword selection for the *entire* training set is superior to random selection. However, a comparison of the results for SPL-seedkey and SPL-allkey shows neither to be consistently superior; in four cases, using keywords for only the initial training set was superior, and in four cases, using keywords for the entire training set was superior.

In summary, the use of a seed set selected using a simple keyword search, composed prior to the review, contributes to the effectiveness of all of the TAR protocols investigated in this study.

7. DISCUSSION

7.1 Random vs. Non-Random Training

The results presented here do not support the commonly advanced position that seed sets, or entire training sets, must be randomly selected [19, 28] [*contra* 11]. Our primary implementation of SPL, in which all training documents were randomly selected, yielded dramatically inferior results to our primary implementations of CAL and SAL, in which none of the training documents were randomly selected. While it is perhaps no surprise to the information retrieval community that active learning generally outperforms random training [22], this result has not previously

been demonstrated for the TAR Problem, and is neither well known nor well accepted within the legal community.

Perhaps more surprising is the fact that a simple keyword search, composed without prior knowledge of the collection, almost always yields a more effective seed set than random selection, whether for CAL, SAL, or SPL. Even when keyword search is used to select *all* training documents, the result is generally superior to that achieved when random selection is used. That said, even if passive learning is enhanced using a keyword-selected seed or training set, it is still dramatically inferior to active learning. It is possible, in theory, that a party could devise keywords that would render passive learning competitive with active learning, but until a formal protocol for constructing such a search can be established, it is impossible to subject the approach to a controlled scientific evaluation. Pending the establishment and scientific validation of such a protocol, reliance on keywords and passive learning remains a questionable practice. On the other hand, the results reported here indicate that it is quite easy for either party (or for the parties together) to construct a keyword search that yields an effective seed set for active learning.

The principal argument in favor of random selection appears to be the concern that non-randomly selected training examples are "less than representative of the entire population of relevant documents" [19, pages 260-261], and therefore might "bias" the learning method, resulting in the exclusion of certain classes of relevant documents. It is easy to imagine that such an effect might occur with SPL; however, it is more difficult to imagine how such a bias could persist through the CAL process.

There are situations in which a finite random sample used as a training set could exclude an identifiable population of relevant documents. By way of example, consider a collection consisting of 1,000,000 emails and 100,000 spreadsheets, of which 10,000 emails and 1,000 spreadsheets were relevant. A random training set consisting of 1,100 documents would contain about 1,000 emails, of which about 10 were relevant, and about 100 spreadsheets, of which, as likely as not, none would be relevant. A machine-learning method might well infer that spreadsheets generally were not relevant, thereby exhibiting a blind spot. Random training tends to be biased in favor of commonly occurring types of relevant documents, at the expense of rare types. Non-random training can counter this bias by uncovering relevant examples of rare types of documents that would be unlikely to appear in a random sample.

7.2 Continuous vs. Simple Active Learning

The differences between the CAL and SAL results arise, we believe, from differences in the design objectives underlying their training methods. The underlying objective of CAL is to find and review as many of the responsive documents as possible, as quickly as possible. The underlying objective of SAL, on the other hand, is to induce the best classifier possible, considering the level of training effort. Generally, the classifier is applied to the collection to produce a review set, which is then subject to manual review.[6] The use of SAL raises the critical issues of (i) what is meant by the "best" classifier, and (ii) how to determine the point at which the best classifier has been achieved (commonly referred to as "stabilization" in the context of TAR). In this study, we arbitrarily define "best" to minimize the total training and review effort necessary to achieve 75% recall, and sidestep the stabilization issue by affording SAL the luxury of an oracle that determines immediately, perfectly, and without cost, when stabilization occurs. In practice, defining and detecting stabilization for SAL (and also for SPL) is "[p]erhaps the most critical question attendant to the use of technology-assisted review for the production of documents" [19, page 263]. In practice, recall and precision of candidate review sets are typically estimated using sampling, and stabilization is deemed to occur when an aggregate measure, such as F_1, appears to be maximized [17]. The choice of a suitable criterion for stabilization, and the cost and uncertainty of sampling to determine when that criterion has been met [3], are fundamental challenges inherent in the use of SAL and SPL that are not addressed in this study; instead, SAL and SPL have been given the benefit of the doubt.

With CAL, each successive classifier is used only to identify – from among those documents not yet reviewed – the next batch of documents for review. How well it would have classified documents that have already been reviewed, how well it would have classified documents beyond the batch selected for review, or how well it would have classified an independent, identically distributed sample of documents, is irrelevant to this purpose. Once it has served this narrow purpose, the classifier is discarded and a new one is created. Because the TAR process continues until as many as possible of the relevant documents are found, the nature of the of documents to which successive classifiers are applied drifts dramatically, as the easy-to-find relevant documents are exhausted and the harder-to-find ones are sought.

For SAL, where training is stopped well before the review is complete, we observed informally that uncertainty sampling was superior to relevance feedback, consistent with previously reported results in machine learning for text categorization [15]. For CAL, our results indicate relevance feedback to be superior.

7.3 When to Terminate the Review

Regardless of the TAR protocol used, the question remains: When to terminate the review? The answer hinges on the proportionality considerations outlined in (U.S.) Federal Rules of Civil Procedure 26(b)(2)(C) and 26(g)(1)(B)(iii), which, respectively, limit discovery if "the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues," and require that discovery be "neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action."

Whether the termination point is determined at stabilization (as for SAL and SPL), or deferred (as for CAL), eventually a legal decision must be made that a reasonable review has been conducted, and that the burden or expense of continuing the review would outweigh the benefit of any additional documents that might be found. The density of responsive documents discovered by CAL appears to fall off monotonically, thus informing the legal decision maker how

[6] In some circumstances – which have not been considered in this study – the review set may be produced to the requesting party without any subsequent review.

much effort would be necessary to find more documents; moreover, Cormack and Mojdeh [7] note that the scores of the responsive documents tend to a normal distribution, and that by fitting such a distribution to the scores, it is possible to estimate recall without resorting to sampling. That said, we leave to future research the issue of how best to determine when to stop.

7.4 Imperfect Training

It has been argued that the accuracy of the human review of the training set is critical, and that a "senior partner" [1, page 184], or even a bi-party committee, should review the training documents [2, page 7]. While the existing scientific literature indicates this concern to be overstated [6, 20, 27], our results further confirm that superior results can be achieved using a single, fallible reviewer. That said, a limitation of our evaluation toolkit is that our simulated reviewer always codes a given document the same way; a real reviewer would be influenced by factors such as the prevalence of responsive documents among those reviewed [23], the order in which the documents were reviewed, and any number of other human factors. We conjecture that these factors would tend to benefit CAL over the other protocols because: (i) the prevalence of responsive documents among those reviewed would be higher, especially at the outset of the review; (ii) similar documents would tend to be reviewed together by virtue of having similar scores; and (iii) the reviewer would gain early insight into the nature of responsive documents without having to wade through a haystack of random or marginal documents looking for an unfamiliar needle. Knowledge of the legally significant documents early in the review process is valuable in its own right. We leave it to future research to confirm or refute our conjecture.

7.5 Limitations

The prevalence of responsive documents in the eight review tasks varies from 0.25% to 3.92%, which is typical for the legal matters with which we have been involved. Others assert that these are examples of "low-prevalence" or "low-richness" collections, for which TAR is unsuitable [19]. We suggest that such assertions may presuppose an SPL protocol [11], which is not as effective on low-prevalence datasets. It may be that SPL methods can achieve better results on higher-prevalence collections (*i.e.*, 10% or more responsive documents). However, no such collections were included in this study because, for the few matters with which we have been involved where the prevalence exceeded 10%, the necessary training and gold-standard assessments were not available. We conjecture that the comparative advantage of CAL over SPL would be decreased, but not eliminated, for high-prevalence collections.

Our evaluation toolkit embodies a number of design choices, the effects of which remain to be explored. Our choices for feature engineering and learning algorithm are state of the art for text classification [5, chapter 11], and we have no indication that another choice would yield materially different results. We reprised most of the experiments in this study using logistic regression, instead of SVM, achieving similar results. A naïve Bayes classifier, on the other hand, achieved generally inferior results overall, but the same relative effectiveness among the protocols. A full exploration of feature engineering and classifier choices remains the subject of future research.

Finally, our use of a batch size of 1,000 was occasioned by efficiency considerations. In each of 100 iterations, we augmented the training set by 1,000 documents, trained the classifier, and scored every document in the collection. Each simulation required several hours of computation; the study required several weeks. For the CAL protocol only, we reran the simulations using a batch size of 100 – entailing ten times as much computation (*i.e.*, several days per simulation) – and achieved slightly better results. The effect of even smaller batch sizes on the effectiveness of TAR protocols remains an open question.

8. CONCLUSION

While the mechanisms and efficacy of active machine learning are well known to the information retrieval community, the legal community has been slow to adopt such technologies, which could help address the growing volume of electronically stored information in (U.S.) legal proceedings. Much of the resistance, we submit, is due to lack of awareness of differences among TAR methods and protocols, and over generalization from one TAR method (typically, a variant of SPL) to all TAR.

Because SPL can be ineffective and inefficient, particularly with the low-prevalence collections that are common in e-discovery, disappointment with such tools may lead lawyers to be reluctant to embrace the use of *all* TAR. Moreover, a number of myths and misconceptions about TAR appear to be closely associated with SPL; notably, that seed and training sets must be randomly selected to avoid "biasing" the learning algorithm.

This study lends no support to the proposition that seed or training sets must be random; to the contrary, keyword seeding, uncertainty sampling, and, in particular, relevance feedback – all non-random methods – improve significantly ($P < 0.01$) upon random sampling.

While active-learning protocols employing uncertainty sampling are clearly more effective than passive-learning protocols, they tend to focus the reviewer's attention on marginal rather than legally significant documents. In addition, uncertainty sampling shares a fundamental weakness with passive learning: the need to define and detect when stabilization has occurred, so as to know when to stop training. In the legal context, this decision is fraught with risk, as premature stabilization could result in insufficient recall and undermine an attorney's certification of having conducted a reasonable search under (U.S.) Federal Rule of Civil Procedure 26(g)(1)(B).

This study highlights an alternative approach – continuous active learning with relevance feedback – that demonstrates superior performance, while avoiding certain problems associated with uncertainty sampling and passive learning. CAL also offers the reviewer the opportunity to quickly identify legally significant documents that can guide litigation strategy, and can readily adapt when new documents are added to the collection, or new issues or interpretations of relevance arise.

There is no reason to presume that the CAL results described here represent the best that can be achieved. Any number of feature engineering methods, learning algorithms, training protocols, and search strategies might yield substantive improvements in the future. The effect of review order and other human factors on training accuracy, and thus overall review effectiveness, may also be substantial.

Nevertheless, the experimental protocol, evaluation toolkit, and results presented here provide a foundation for further studies to investigate these and other possible approaches to improve the state of the art in TAR for e-discovery.

9. ACKNOWLEDGEMENT

Cormack's research is supported by a Discovery grant and a Research Tools and Instruments grant from the Natural Sciences and Engineering Research Council of Canada.

10. REFERENCES

[1] *Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, S.D.N.Y., 2012.

[2] Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI"), *In Re: Actos (Pioglitazone) Products Liability Litigation*, MDL No. 6:11-md-2299, W.D. La., July 27, 2012.

[3] M. Bagdouri, W. Webber, D. D. Lewis, and D. W. Oard. Towards minimizing the annotation cost of certified text classification. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 989–998, 2013.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, 2008.

[5] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[6] J. Cheng, A. Jones, C. Privault, and J.-M. Renders. Soft labeling for multi-pass document review. *ICAIL 2013 DESI V Workshop*, 2013.

[7] G. V. Cormack and M. Mojdeh. Machine learning for information retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks. *The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

[8] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3):1–48, 2011.

[9] M. R. Grossman and G. V. Cormack. Inconsistent responsiveness determination in document review: Difference of opinion or human error? *Pace Law Review*, 32(2):267–288, 2012.

[10] M. R. Grossman and G. V. Cormack. The Grossman-Cormack glossary of technology-assisted review with foreword by John M. Facciola, U.S. Magistrate Judge. *Federal Courts Law Review*, 7(1):1–34, 2013.

[11] M. R. Grossman and G. V. Cormack. Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review." *Federal Courts Law Review*, 1, to appear 2014.

[12] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 Legal Track. *The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

[13] C. Hogan, J. Reinhart, D. Brassil, M. Gerber, S. Rugani, and T. Jade. H5 at TREC 2008 Legal Interactive: User modeling, assessment & measurement. *The Seventeenth Text REtrieval Conference (TREC 2008)*, 2008.

[14] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[15] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[16] D. W. Oard and W. Webber. Information retrieval for e-discovery. *Information Retrieval*, 6(1):1–140, 2012.

[17] Y. Ravid. *System for Enhancing Expert-Based Computerized Analysis of a Set of Digital Documents and Methods Useful in Conjunction Therewith*. United States Patent 8527523, 2013.

[18] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[19] K. Schieneman and T. Gricks. The implications of Rule 26(g) on the use of technology-assisted review. *Federal Courts Law Review*, 7(1):239–274, 2013.

[20] J. C. Scholtes, T. van Cann, and M. Mack. The impact of incorrect training sets and rolling collections on technology-assisted review. *ICAIL 2013 DESI V Workshop*, 2013.

[21] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[22] B. Settles. *Active learning literature survey*. University of Wisconsin, Madison, 2010.

[23] M. D. Smucker and C. P. Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602, 2010.

[24] S. Tomlinson. Learning Task experiments in the TREC 2010 Legal Track. *The Nineteenth Text REtrieval Conference (TREC 2010)*, 2010.

[25] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[26] W. Webber, D. W. Oard, F. Scholer, and B. Hedin. Assessor error in stratified evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 623–632, 2010.

[27] W. Webber and J. Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 929–932, 2013.

[28] C. Yablon and N. Landsman-Roos. Predictive coding: Emerging questions and concerns. *South Carolina Law Review*, 64(3):633–765, 2013.

EXHIBIT B

THE FEDERAL COURTS LAW REVIEW

Volume 7, Issue 1 2014

Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review"

Maura R. Grossman[†] and
Gordon V. Cormack[††]

ABSTRACT

Approaches to technology-assisted review ("TAR") and its validation—presented as "obligations" under Federal Rule 26(g) in a recent article by Karl Schieneman and Thomas C. Gricks III—could, if prescribed, impair the effectiveness of TAR and increase its expense, thereby compromising a primary objective of Rule 26(g): to ensure a reasonable production at a proportionate cost.

Extraordinary culling efforts to enrich the collection are likely to eliminate large amounts of responsive information, while affording only the illusion of improved review or validation effectiveness. In addition, empirical evidence shows that the use of random selection for seed or training sets is inferior to keyword selection, and substantially inferior to the use of active learning—non-random selection determined by a machine-learning algorithm. Finally, exclusive focus on a particular statistical test, applied to a single phase of a review effort, does not provide adequate assurance of a reasonable production, and may be unduly burdensome. Validation should consider all available evidence concerning the effectiveness of the end-to-end review process, including prior scientific evaluation of the TAR method, its proper application by qualified individuals, and proportionate *post hoc* sampling for confirmation purposes.

TABLE OF CONTENTS

† A.B., Brown University; M.A. and Ph.D. in Psychology, Gordon F. Derner Institute of Advanced Psychological Studies, Adelphi University; J.D., Georgetown University Law Center; Of Counsel, Wachtell, Lipton, Rosen & Katz. The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

†† B.Sc., M.Sc., and Ph.D. in Computer Science, University of Manitoba; Professor and Co-Director of the Information Retrieval Group, David R. Cheriton School of Computer Science, University of Waterloo.

I. INTRODUCTION

In a recent Federal Courts Law Review article titled "The Implications of Rule 26(g) on the Use of Technology-Assisted Review,"[1] Karl Schieneman and Thomas C. Gricks III address important issues surrounding the use of technology-assisted review ("TAR") in electronic discovery. Schieneman and Gricks assert that Federal Rule of Civil Procedure 26(g) ("Rule 26(g)") imposes "unique obligations"[2] on responding parties that use TAR to produce documents in response to discovery requests. These obligations include:

1. Exercising "greater care in the collection of ESI" to "optimize and manage the richness of the database," by narrowing or culling the document collection "in order to [maximize] the fraction of relevant documents processed into the tool";[3] and

2. Training the TAR system using a random "seed" or "training" set, as opposed to one relying on judgmental sampling, which "may not be representative of the entire population of electronic documents within a given collection."[4]

Schieneman and Gricks' assertion appears to be driven by their premise that the "goal of technology-assisted review is to achieve an acceptable level of recall and precision, in light of proportionality considerations, based on a statistical quality control analysis."[5] They

1. Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 239 (2013), *available at* http://www.fclr.org/fclr/articles/html/2010/Gricks.pdf (hereinafter "Schieneman & Gricks").

2. *Id.* at 240.

3. *Id.* at 249, 251–52. "ESI" refers to "Electronically Stored Information." Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review with Foreword by John M. Facciola, U.S. Magistrate Judge*, 7 Fed. Cts. L. Rev. 1 (2013), at 15, 16, *available at* http://www.fclr.org/fclr/articles/html/2010/grossman.pdf (hereinafter "*TAR Glossary*").

4. Schieneman & Gricks, *supra* note 1, at 260.

5. *Id.* at 269. "Recall" is defined as "The fraction of Relevant Documents that are identified as Relevant by a search or review effort." *TAR Glossary*, *supra* note 3, at 27. "Precision" is defined

state that, to demonstrate "reasonable inquiry" under Rule 26(g), it is necessary to conduct random sampling sufficient to estimate recall and precision with a margin of error of at most ±5%, and a confidence level of at least 95%.[6]

We submit that the goal of technology-assisted review, or any other review,[7] is to identify for production as much responsive information as reasonably possible, at a proportionate cost. While statistics such as recall and precision are *measures* of success in achieving that goal, they are not the goal in itself. Indeed, when statistical outcomes become goals, they are subject to Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."[8] Schieneman and Gricks' asserted obligations would have the statistical tail wagging the best-practices dog; that is, they would have the TAR process dictated by sampling practices rather than by what would achieve the best possible review.

We are concerned that Schieneman and Gricks' prescriptions may offer false assurance that by rote adherence to a particular formulaic approach, a responding party can meet the "reasonable inquiry" requirement of Rule 26(g). Adoption of their prescriptions could also invite a requesting party to demand evidence of adherence "at every step of the process,"[9] leading to discovery about discovery and successive motions practice. Moreover, in many circumstances, rigid adherence to Schieneman and Gricks' proposals would incur disproportionate burden and cost, while compromising the effectiveness of—or even precluding—perfectly reasonable approaches to TAR and to validation.

Schieneman and Gricks appear to presuppose a particular approach to TAR, which we refer to here as "Simple Passive Learning" or "SPL," and further, to conflate this approach with a particular approach to validation. But there is no single approach to TAR or to validation, nor any necessity that TAR and validation employ a common approach.

We use, as an example of a TAR method that does not fit Schieneman and Gricks' assumptions, the method employed by Gordon V. Cormack and Mona Mojdeh in the course of their participation in the

---

(footnote continued)

as "The fraction of Documents identified as Relevant by a search or review effort that are in fact Relevant." *Id.* at 25.

6. Schieneman & Gricks, *supra* note 1, at 270 ("Regardless of the specific measure being evaluated, utilizing a sample having a confidence level of either 95% or 99%, and a nominal confidence interval of between ±2% and ±5% will satisfy the reasonable inquiry requirements of Rule 26(g).").

7. While we focus our comments on TAR, we dispute the logic by which Schieneman and Gricks conclude that the purpose, conduct, or validation of TAR methods differs from that of traditional review methods so as to incur "unique obligations."

8. Marilyn Strathern, *'Improving Ratings': Audit in the British University System*, 5 EUR. REV. 305, 308 (1997), *available at* http://journals.cambridge.org/article_S1062798700002660.

9. Schieneman & Gricks, *supra* note 1, at 248 ("[A]dherence to the principles of Rule 26(g) is critical at every step of the process when technology-assisted review is used. . . ." *See also id.* at 247 ("Ultimately, whether counsel has adequately discharged applicable Rule 26(g) obligations will typically be 'a fact intensive inquiry that requires evaluation of the procedures the producing party adopted during discovery. . . .'" (footnote omitted)).

TREC 2009 Legal Track,[10] and subsequently employed by the authors in fifty civil and regulatory matters. We refer to this method as "Continuous Active Learning" or "CAL."

We argue that the use of TAR—whether employing SPL or CAL—does not occasion the use of extraordinary efforts to "optimize and manage [] richness," and that such efforts may compromise not only the quality and cost effectiveness of the review, but also the accuracy and cost effectiveness of validation. We also argue that the use of TAR does not require the use of random seed or training sets, and that the removal of judgmental and other non-random input may impair the quality and increase the cost of the review. Finally, we argue that the proposed validation method—setting a recall target during "stabilization" and sampling the "review set" to ensure that the target is met[11]—incorrectly focuses on an intermediate phase of the end-to-end review process—a phase that may be a necessary part of SPL, but is absent from CAL. Validation, we argue, is best achieved by considering the end-to-end effectiveness of the review, and evaluating the totality of the evidence derived from multiple sources, not by considering only a single target measure applied to a particular phase of the review process.

We illustrate our arguments by considering the application of alternative TAR and validation approaches within the context of a hypothetical matter:

> The responding party has employed customary practices to identify custodians and ESI sources that may contain responsive information. After de-NISTing,[12] deduplication, and date restriction, one million documents that meet the criteria have been imported into a review tool. Unbeknownst to the responding party, the collection contains ten thousand responsive documents. The goal of the review is to find as many of these responsive documents as possible, at a proportionate cost, while ensuring through reasonable inquiry that this goal has been met.

---

10. Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS (2009), at 2–6, *available at* http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf (hereinafter "Cormack & Mojdeh"). *See also* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), at 31–34, *available at* http://jolt.richmond.edu/v17i3/article11.pdf.

11. *See* Schieneman & Gricks, *supra* note 1, at 263–73.

12. "De-NIST[ing]" is defined as "The use of an automated filter program that screens files against the NIST list in order to remove files that are generally accepted to be system generated and have no substantive value in most instances." The Sedona Conference, *The Sedona Conference Glossary: E-Discovery & Digital Information Management* (4th ed. 2014), at 13, *available at* https://thesedonaconference.org/publication/The Sedona Conference%C2%AE Glossary.

The remainder of this paper is organized as follows. Section II presents CAL and the evidence that it works at least as well as any TAR method reported in the scientific literature. Section III describes SPL and contrasts it with CAL. Section IV argues that extraordinary efforts to "optimize and manage the richness of the database" provide only the illusion of improved recall and easier validation; an apples-to-apples comparison reveals that such efforts are likely to reduce the number of responsive documents found by the review, without any commensurate reduction of "uncertainty in the result."[13] Section V presents evidence that modifying CAL to use random training examples actually harms its effectiveness, and that the use of judgmental training examples can improve SPL. Section VI discusses the need to validate the end-to-end effectiveness of the entire review effort, not just the specific portion of the review process that Schieneman and Gricks refer to as "technology-assisted review."[14] Section VII discusses the limitations of recall as a validation measure, the statistical and methodological challenges of measuring recall, and the mathematical unsoundness of the various sampling methods advanced by Schieneman and Gricks, as well the closely related "eRecall" method.[15] In Section VIII, we offer some suggestions for a way forward. Relying on the elements of the *Daubert* test, we argue that the principal focus of validation should be on (i) prior scientific evaluation of the TAR method, (ii) ensuring its proper application by qualified individuals, and (iii) proportionate *post hoc* sampling for confirmation purposes. Section IX offers our conclusions.

II. CONTINUOUS ACTIVE LEARNING

The TAR method found by the authors to be most effective is Continuous Active Learning ("CAL"). CAL involves two interactive tools: a keyword search system with relevance ranking,[16] and a machine-learning algorithm.[17] At the outset of the TAR process, keyword searches are performed and some of the top-ranked documents from each search are coded by a human reviewer as responsive or not. These coded documents (the "seed set") are used to train the learning algorithm, which ranks each document in the collection by the likelihood that it

13. Schieneman & Gricks, *supra* note 1, at 251–52.

14. *Id.* at 269.

15. Herbert L. Roitblat, *A Tutorial on Sampling in Predictive Coding* (OrcaTec LLC 2013), at 3, *available at* http://orcatec.com/2013/10/07/a-tutorial-on-sampling-in-predictive-coding/ ("[W]e call Recall estimated from Elusion and Prevalence eRecall to distinguish it from Recall computed directly. . . ."); Herbert L. Roitblat, *Measurement in eDiscovery: A Technical White Paper* (OrcaTec LLC 2013), at 10, *available at* http://orcatec.com/2013/10/06/measurement-in-ediscovery-a-technical-white-paper/ (hereinafter "*Measurement in eDiscovery*") ("Estimating Recall from Elusion can be called eRecall.").

16. "Relevance Ranking" is defined as "A search method in which the results are ranked from the most likely to the least likely to be Relevant to an Information Need; the result of such ranking. Google Web Search is an example of Relevance Ranking." *TAR Glossary*, *supra* note 3, at 28.

17. "Machine Learning" is defined as "The use of a computer Algorithm to organize or Classify Documents by analyzing their Features." *Id.* at 22.

contains responsive information. The top-ranked documents that have not yet been coded are then coded by a human reviewer, and the learning algorithm is retrained using all coded documents. This process of training and coding is repeated until the number of top-ranked documents containing responsive information drops precipitously. At that time, quality-assurance efforts, including, but not limited to, additional keyword searches, score modeling,[18] and random sampling, are undertaken. If any of these efforts uncovers more documents containing responsive information, the CAL process is restarted, provided that the quantity, novelty, and importance of the newly identified information justify the additional effort.

CAL was shown, in what Schieneman and Gricks refer to as the "preeminent study of the effectiveness of information retrieval techniques in the legal field"[19] (the "JOLT Study"[20]), to be superior to manual review for four review tasks conducted at TREC 2009.[21] Yet CAL would fail Schieneman and Gricks' Rule 26(g) test. In the four tasks reported in the JOLT Study, Cormack and Mojdeh achieved superior results applying CAL directly, without "optimiz[ing] and manag[ing] the richness" of the collections, even though each collection had "low richness," as defined by Schieneman and Gricks[22] (*i.e.*, 0.3%, 0.7%, 0.3%, and 1.5%).[23] Of the matters in which the authors have subsequently applied CAL—without "optimiz[ing] and manag[ing] [] richness"—more than 90% have involved collections with "low richness."[24] Moreover, at TREC 2009, and in all of the authors' subsequent matters, keyword search (*i.e.*, "judgmental selection"[25]) was used to create the seed set.

---

18. *See* Cormack & Mojdeh, *supra* note 10, at 7 (Figure 4 and accompanying text).

19. Schieneman & Gricks, *supra* note 1, at 263 ("The starting point for the analysis of technology-assisted review is the consideration of the relative effectiveness of available alternatives. The preeminent study of the effectiveness of information retrieval techniques in the legal field was the JOLT Study, published in 2011.").

20. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), *available at* http://jolt.richmond.edu/v17i3/article11.pdf.

21. Bruce Hedin, Stephen Tomlinson, Jason R. Baron & Douglas W. Oard, *Overview of the TREC 2009 Legal Track*, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS, at 17 Table 6 (Topics 201, 202, 203, and 207), *available at* http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf (hereinafter, "Hedin et al."). It is worth noting that e-discovery service provider H5, using a radically different, rule-based method—which also does not appear to conform to Schieneman and Gricks' asserted Rule 26(g) obligations—did equally well on a fifth review task at TREC 2009. *Id.* (Topic 204).

22. Schieneman & Gricks, *supra* note 1, at 250 ("a prevalence of only 1% (a low prevalence) . . . .").

23. Hedin et al., *supra* note 21, at 16 Table 5 (Topics 201, 202, 203, and 207).

24. Our median richness has been less than 1%; we have had very few matters with richness above 3%. *Compare with* Schieneman & Gricks, *supra* note 1, at 249 ("Richness tends to be between five percent (5%) and ten percent (10%) of the total collection, but may be greater or even an order of magnitude less," citing *Measurement in eDiscovery*, *supra* note 15, at 6 ("We tend to see that around 5 [sic] 5-10% of the documents in an average collection are responsive, though we have seen higher (up to around 50%) and lower (around 0.5%) Prevalence on occasion. Others report even lower Prevalence.")).

25. Schieneman & Gricks, *supra* note 1, at 259.

The hypothetical matter represents a situation that would be typical for the authors' application of TAR. The richness of the collection is 1% (10,000 responsive documents in a collection of 1,000,000 documents). We would expect to find representative exemplars within several hours of searching, and then employ an active-learning algorithm until most of the responsive documents were found. On the order of 20,000 documents would need to be reviewed before the bulk of the responsive documents were identified.[26] We would know that our approach was reasonably likely to succeed, because we had employed the same technique fifty times to a variety of collections and production requests, always achieving a successful result. During the conduct of the review, internal measures would reinforce our expectations, and after the review, sampling would further corroborate our belief that a reasonable review had been conducted.

III. SIMPLE PASSIVE LEARNING

The TAR method that appears to be assumed by Schieneman and Gricks, and many others, is Simple Passive Learning ("SPL"). A review involving SPL has two distinct phases: training and review. In the training phase, a set of training documents (often referred to as the "seed set") is identified and coded as responsive or not. At the end of the training phase, the coded training documents are used as input to a learning algorithm, which either identifies a subset of the collection as likely responsive (the "review set"), or scores each document in the collection by the likelihood—as estimated by the learning algorithm—that it is responsive, in which case the review set consists of the documents achieving a score equal to or exceeding some "threshold" or "cutoff" value. In the review phase, the review set is coded manually, and only those documents coded as responsive (less any coded as privileged) are produced. On rare occasions, the review phase may be omitted, and the review set produced without further review.

SPL methods diverge from one another in how the training documents are selected. The principal approaches are judgmental selection (*e.g.*, keyword search), random selection, or a combination of both. Judgmental selection relies on the skill of the searcher to find appropriate training examples; random selection relies on chance; a combination hedges its bets.

SPL differs from CAL in two important respects: In SPL, the learning algorithm plays no role in the selection of training examples, and the review phase is separate and distinct from the training phase. As a consequence, the size and composition of the training set play a critical role in SPL; once these choices are made and the training set is coded, the review set is fixed, and the quality of the review is largely

26. In our experience, it is generally necessary to review about twice as many documents as the number of responsive documents in the collection.

predetermined. In SPL, the critical step of determining whether a particular training set (and hence review set) is adequate, or whether it could be enhanced sufficiently to justify additional training effort, is known as "stabilization."[27]

To achieve stabilization, SPL methods typically employ an iterative training phase in which a candidate training set is constructed (using one of the possible selection methods) and used by the learning algorithm to create a candidate review set. If sampling indicates that the review set is "adequate," the training phase is complete; otherwise, the training set is revised (typically by adding more examples), a new candidate review set is created, and sampling is repeated until stabilization is achieved.

A precise definition of an "adequate review set" remains elusive, as does the design of a sampling strategy to determine adequacy, however defined. As suggested by Schieneman and Gricks, one might specify target levels of precision and recall, informed by what could reasonably be achieved with proportionate effort.[28] Unfortunately, determining what could reasonably be achieved involves predicting the future: How much could the review set be improved, for a given amount of additional training effort? If the targets are set too high, stabilization might never occur; if the targets are set too low, stabilization might occur too early, resulting in an inferior review compared to what could have been achieved with reasonable additional effort.

In our hypothetical matter, target levels of 75% might be set for both precision and recall. We might begin with a seed set of 1,000 documents identified using random selection, keyword search, or a combination of both. These documents would be used to train the learning algorithm and to identify a candidate review set. A random sample of 1,000 documents might then be drawn from the collection and used to estimate the precision and recall of the review set. Assuming the estimated precision or recall were inadequate (*i.e.*, < 75%), the sample would be added to the training set, and the process repeated, until both the estimated precision and recall exceeded 75%. At this point, the

---

27. *Id.* at 263. Schieneman and Gricks do not define the term "stabilization" in their article. Stabilization is a vendor term that has been used to refer to the point at which further training will not improve the effectiveness of the learning algorithm. *See, e.g.*, Chris Dale, *Far From the Black Box: Explaining Equivio Relevance to Lawyers – White Paper* (Equivio 2012), at 9, *available at* http://www.equivio.com/files/files/White_Paper_–_Far_from_the_Black_Box:_Explaining_Equivio Relevance_to_Lawyers.pdf ("[T]here comes a point when further training adds nothing to the system's understanding. This is known in Equivio Relevance as the 'stabilisation point'. . . ."). *See also Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, 187 (S.D.N.Y. 2012) (discussing stabilization in the context of "the training of the [TAR] software").

28. Schieneman & Gricks, *supra* note 1, at 266 ("Ultimately, what constitutes an acceptable level of recall will be based on the totality of the circumstances and, absent agreement, counsel should be prepared to make a proper showing under Federal Rule 26(b)(2)(C) to support any proposed target or actual result."). *See also id.* at 263 ("Perhaps the most critical question attendant to the use of technology-assisted review for the production of documents is this: what levels of recall and precision are sufficient under Federal Rule of Civil Procedure 26(g)? Not surprisingly, given the incipience of technology-assisted review as a document review and production technique, neither the case law nor the Federal Rules provide a bright-line answer."), 267 ("[T]he party proposing the limitation must demonstrate the need to limit that discovery (*i.e.*, establishing a recall objective)." (footnote omitted)).

training phase would be deemed complete, and the review phase would begin, using the final candidate review set.

Although not envisioned in the TAR process described by Schieneman and Gricks, sampling following the review phase could be used to confirm the expectation that a reasonable review had been conducted. *See* Section VI.

IV. COLLECTION ENRICHMENT

As illustrated in this section, an obligation to enrich the collection would increase the complexity and cost of the review, while providing only the illusion of improved validation. At the start of the review, it would be necessary for the responding party to sample the collection to estimate richness, so as to determine how much, if any, enrichment was required. In order to enrich the collection, it would be necessary to discard the vast majority of the documents, at least some of which would almost certainly be responsive and, by virtue of being discarded prior to the TAR effort, never reviewed or produced. To verify that a negligible number of responsive documents were discarded through enrichment, it would be necessary to sample the discarded documents—an even more challenging task than sampling the original collection, because the discard pile would have substantially lower richness. Moreover, target measures derived from the enriched collection would not be the same as—indeed, would be incomparable to—target measures derived from the original collection. As a consequence, it is entirely possible, as illustrated below, that a review effort achieving a higher recall on the enriched collection might actually find fewer responsive documents overall—and incur a higher level of uncertainty as to the quality of the result—than a review effort achieving a lower recall on the original collection. In short, collection enrichment simply moves the statistical goalposts.

In our hypothetical matter, a random sample would be used to estimate the richness of the original collection in order to determine how much enrichment would be necessary. Determining the sample size for this estimate would not be a trivial matter, as the choice would require a guess as to the richness, which, of course, is not known prior to taking the sample. The choice would further require an assessment of the responding party's tolerance for error in the estimate. An inadequate sample size could result in the estimate being either too low or too high: Too low an estimate would mislead the responding party to undertake excessive enrichment, while too high an estimate would mislead the responding party to undertake insufficient enrichment. A full discussion of sampling for this purpose is beyond the scope of this paper; for the purpose of illustration, we will arbitrarily choose a sample size of N=1,000. For this value of N, given the fact that (unbeknownst to the responding party) 10,000 of the 1,000,000 documents are responsive, approximately 10—but probably not exactly 10—of the documents in the sample would be responsive. For simplicity, let us assume that the

responding party were lucky and the number were exactly 10. The point estimate for richness would therefore be 10/1,000, or 1%.

In order to increase richness to 10%—a tenfold increase—it would be necessary to reduce the size of the collection at least tenfold; that is, to discard at least nine-tenths of the collection, or 900,000 documents. Assuming, for simplicity, that all of the discarded documents were truly non-responsive (*i.e.*, that *none* of the 10,000 responsive documents were among those discarded), the enriched collection would then have 100,000 documents, of which 10,000 were responsive (*i.e.*, 10% richness). This assumption, however, is not realistic; in reality, the discard pile would contain at least *some* percentage of responsive documents. Suppose that percentage were only one-third of 1% (0.33%). In that case, the enrichment process would discard 3,000 responsive documents and would yield an enriched collection with 100,000 documents, of which 7,000 were responsive (*i.e.*, 7% richness). By the same token, if, instead, two-thirds of 1% (0.67%) of the discard pile were responsive, 6,000 responsive documents would be lost, and the resulting enriched collection would have 4,000 responsive documents (*i.e.*, 4% richness). Clearly, it is important to know how many of the discarded documents are responsive, not only to achieve the desired richness, but, more importantly, to ensure that the baby has not been thrown out with the bathwater. Estimating the number of discarded responsive documents is fraught with exactly the same statistical peril—low richness—that optimizing and managing richness is intended to avoid.

Suppose, in our hypothetical matter, the richness of the enriched collection was 7% (*i.e.*, the enriched collection contained 7,000 responsive documents). Suppose further that 7% richness, although falling short of the 10% target, was deemed to be an adequate richness to proceed with TAR. If a TAR method were then applied to the enriched collection, achieving 80% recall on that collection, the result would be that 80% of the 7,000 responsive documents (or 5,600 responsive documents) would be identified by the enrichment and review processes combined. In contrast, if the same TAR method were applied directly to the original collection, achieving only 70% recall—instead of 80% recall—the result would be to find 70% of 10,000 documents, or 7,000 responsive documents. In short, the enrichment process offers the illusion of better recall (80% as compared to 70%), while actually identifying fewer responsive documents (5,600 as compared to 7,000). The comparison of the two recall values is specious, as they are derived from different collections. A valid comparison demands that both recall values be derived from the original collection, in which case we see that, in this simple example, the recall of TAR alone is 70%, while the recall of enrichment plus TAR is 56%.

We posit that the impetus for Schieneman and Gricks' advocacy of collection enrichment is the presumption that all TAR methods use SPL and, in particular, require precise statistical estimates to achieve

stabilization.[29] Only in light of this presumption can we make sense of Schieneman and Gricks' assertions that (i) "technology-assisted review depends primarily upon statistics to validate effectiveness";[30] (ii) "Rule 26(g) requires counsel to recognize, understand, and endeavor to avoid high levels of uncertainty, and minimize the margin of error, in their use of technology-assisted review";[31] and (iii) "these obligations require counsel to understand the impact of richness on the ability of technology-assisted review to cull relevant documents from the database, and to develop a strategy for avoiding as much uncertainty in the result as is reasonably possible."[32] As noted in Section II, CAL consistently achieves high recall from low-prevalence collections, and *does not* in any way depend on using statistics yielding a low margin of error. In contrast, as outlined in Section III, SPL *does* use statistics to determine when stabilization has occurred, so as to terminate the training phase. A high level of uncertainty could, in this situation, result in a premature end to training, an inadequate review set, and ultimately, an inadequate production. If an SPL method—or any TAR method—does not work well for collections with low prevalence, it should not be applied to collections with low prevalence. Extraordinary culling effort for the purpose of enrichment, in ways "not previously associated with traditional review techniques,"[33] is not the answer.

V. RANDOM TRAINING

It is difficult to determine precisely which training regimen(s) Schieneman and Gricks advocate (or discourage) when they discuss "the manner in which the training set or seed set of documents is generated, *i.e.*, through judgmental selection *or* random selection."[34] The choice between judgmental and random selection is a false dichotomy because neither method must be used to the exclusion of the other, and because there are other methods—such as active learning—that are neither random nor judgmental. It is possible to read Schieneman and Gricks as (i) requiring exclusively random selection (and therefore prohibiting both judgmental and other non-random selection methods), or (ii) permitting judgmental selection or other non-random methods only if used in conjunction with random selection.

---

29. The most common impetus for collection enrichment (*i.e.*, culling) is to reduce the costs arising from volume-based vendor pricing, not to improve statistical estimates. *See*, *e.g.*, Biomet's Submission in Support of Its Discovery Efforts at 19, *In Re: Biomet M2a Magnum Implants Prods. Liab. Litig.*, No. 3:12-md-2391 (RLM) (CAN) (N.D. Ind. Apr. 4, 2013) ("Biomet has already spent over $1 million on conducting predictive coding on the 2.5+ million documents selected by search terms. Expanding predictive coding to the remaining 15.5+ million documents would cost between $2 and $3.25 million more in processing costs. . . ." (citations omitted)).

30. Schieneman & Gricks, *supra* note 1, at 249.

31. *Id.* at 251.

32. *Id.*

33. *Id.* at 249. *See also id.* ("Rule 26(g) requires counsel to exercise greater care in the collection of ESI, in order to optimize the fraction of relevant documents processed into the tool.").

34. *Id.* at 259 (emphasis added).

Regardless of which interpretation is intended, Schieneman and Gricks' thesis is predicated on the belief that non-random training regimens use training documents that may be "less than representative of the entire population of relevant documents," and therefore "run afoul of Rule 26(g)."[35] Schieneman and Gricks posit further that "[o]ne of the ways to avoid disagreement over the methodology used to train the tool" is "to share the training set with opposing counsel."[36] However, they assume, without evidence, that (i) a suitable seed[37] or training set for TAR must "reflect[]" the "full spectrum of relevant documents";[38] (ii) a random seed or training set is necessarily suitable (whereas a judgmentally selected set is not);[39] and (iii) it is possible for the requesting party to discern from the contents of a seed or training set whether or not it is "representative of the entire population of relevant documents," or otherwise suitable for training.[40]

The notion that seed or training sets must be random appears to derive from a false equivalence between random sampling for the purpose of *statistical estimation* and the random selection of examples for the purpose of *training a learning algorithm*.[41] It does not follow that because random sampling is necessary for statistical estimation, it is also necessary for training a learning algorithm. The argument to that effect is akin to saying that because a hammer is the proper tool to drive in a nail, it should also be used to pound in a screw.

A recent experimental study by the authors "lends no support to the proposition that seed or training sets must be random; to the contrary, keyword seeding, uncertainty sampling,[42] and, in particular, relevance feedback[43]—all non-random methods—improve significantly [at the 99% confidence level] upon random sampling."[44] Specifically, CAL—in

---

35. *Id.* at 260–61.
36. *Id.* at 261.
37. Although acknowledging ambiguity in the meaning of the phrase "seed set," Schieneman and Gricks never define the sense in which they use the phrase, and gloss over the distinction between seed set and training set. *See id.* at 259–61. As illustrated in Sections II and III, the seed set constitutes only a small fraction of the training set when using CAL, whereas the seed set is often taken to be the entire training set when using SPL. Schieneman and Gricks' conflation of seed set and training set is consistent with our impression that, in their arguments, they have considered only SPL, overlooking CAL and potentially other TAR methods.
38. *Id.* at 261.
39. *Id.* at 260–61.
40. *Id.*
41. The passage from the *TAR Glossary* defining "Judgmental Sampling" that is quoted by Schieneman and Gricks (*supra* note 1, at 260), concerns *statistical estimation*, not machine learning. *See TAR Glossary*, *supra* note 3, at 21 ("Unlike a Random Sample, the *statistical properties* of a Judgmental Sample may not be extrapolated to the entire Population." (emphasis added)).
42. "Uncertainty Sampling" is defined as "An Active Learning approach in which the Machine Learning Algorithm selects the Documents as to which it is least certain about Relevance, for Coding by the Subject Matter Expert(s), and addition to the Training Set." *TAR Glossary*, *supra* note 3, at 33-34.
43. "Relevance Feedback" is defined as "An Active Learning process in which the Documents with the highest likelihood of Relevance are coded by a human, and added to the Training Set." *Id.* at 28.
44. Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, in PROCEEDINGS OF THE 37TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION

which *none* of the training documents were randomly selected—substantially outperformed SPL—in which *all* of the training documents were randomly selected.[45] On every one of eight review tasks, the CAL approach—using keyword seeding and active learning—found more responsive documents, with less review effort, than SPL.[46] This result is partially illustrated in Table 1, which shows, for CAL and SPL, the total review effort required to achieve a recall of 75%.[47]

Table 1: Review Effort to Achieve 75% Recall

	CAL	SPL		
Matter	**Total Effort**	**Training Phase Effort**	**Review Phase Effort**	**Total Effort**
201	**6,000**	44,000	12,000	**56,000**
202	**11,000**	7,000	19,000	**26,000**
203	**6,000**	9,000	90,000	**99,000**
207	**11,000**	9,000	26,000	**35,000**
A	**11,000**	27,000	58,000	**85,000**
B	**8,000**	7,000	13,000	**20,000**
C	**4,000**	3,000	4,000	**7,000**
D	**18,000**	6,000	31,000	**37,000**

Total Effort is measured in terms of the number of documents reviewed to achieve the target recall of 75%. For CAL, Total Effort includes both training and review, which are not separate phases. For SPL, training and review are distinct phases; thus, Total Effort reflects the sum of both.

Through testing variants of CAL and SPL, the study also determined that "a simple keyword search, composed prior to the review,

(footnote continued)

RETRIEVAL (SIGIR '14) (July 2014), at 9, authors' copy *available at* http://cormack.uwaterloo.ca/cormack/calstudy/ (hereinafter "Cormack & Grossman").

45. *Id.* at 4, 5 Figure 1, 7 Table 5.

46. *Id.*

47. The Total Effort data reflected in columns 2 and 5 of Table 1 are taken from Cormack & Grossman, *supra* note 44, at 7 Table 5. The Training and Review Phase Effort data reflected in columns 3 and 4 of Table 1 are on file with the authors. For SPL, the optimal training-set size was chosen with the benefit of hindsight: The experiment was repeated with 100 training-set sizes and the result yielding the *least* total effort is reported in Table 1. In practice, the optimal training-set size would not be known in advance; it would be necessary to achieve stabilization (as described in Section III), entailing effort and uncertainty. The results reported here give SPL the benefit of the doubt, assuming that stabilization occurred at exactly the optimal training-set size, and that no extra documents were reviewed to achieve stabilization.

contributes to the effectiveness of all of the TAR protocols investigated in this study."[48]

These results call into question Schieneman and Gricks' thesis, however interpreted. For CAL, judgmental selection of the initial training set (the seed set) "generally yields superior results to random selection."[49] Active learning—a non-random method—works better than random selection for the rest of the training examples.[50] For SPL, using *some* judgmentally selected training examples improves on using entirely randomly selected ones, while using *entirely* judgmentally selected training examples often improves on using entirely randomly selected ones.[51] Overall, judgmental selection and active learning combined—without any random selection—work best of all.[52]

Our results are not entirely inconsistent with one premise behind Schieneman and Gricks' thesis: Using judgmental sampling *alone* may be a risky proposition. Indeed, the widely held belief—based on experience with some TAR tools that use SPL—that the judgmental seed set constitutes the *only* training input to the TAR tool appears to be the source of much of the angst that has been expressed regarding the use of all TAR tools, even those that also use active learning. Perhaps a gifted searcher (a "TAR Whisperer") could judgmentally select precisely the right set of examples to correctly train the SPL tool, but can the typical responding party be relied upon to conduct an adequate search? Accordingly, the argument goes, all human judgment should be eschewed in favor of random sampling. While the fear of *exclusively* judgmental selection may be valid, the proposed remedy is not.

Schieneman and Gricks' suggestion, shared by others, that "[o]ne of the ways to avoid disagreement over the methodology used to train the tool" is "to share the training set with opposing counsel,"[53] is, we believe, ill-founded. Disclosure of the seed or training set offers false comfort to the requesting party, in much the same way that disclosure of the keywords to be used for culling—without testing them—provides limited information. An examination of the responsive training documents would reveal inadequacies only to the extent that the requesting party was aware of the "full spectrum of relevant documents"[54] in the collection. But if the requesting party had such knowledge, it could more fruitfully examine the production itself, as opposed to the training set. More importantly, however, alleged "gaps" in the training set do not necessarily translate into gaps in the production, as an effective active-learning tool should readily fill those gaps.[55]

---

48. Cormack & Grossman, *supra* note 44, at 7.
49. *Id.*
50. *Id.* at 1.
51. *Id.* at 7, 8.
52. *Id.* at 4, 5 Figure 1, 7.
53. Schieneman & Gricks, *supra* note 1, at 251.
54. *Id.* at 261.
55. As evidenced by high levels of recall achieved across multiple review tasks, CAL was able to uncover documents representative of the "full spectrum of relevant documents" using a seed

Examination of the non-responsive training documents offers even less probative value than examination of the responsive training documents. Without access to the TAR tool and the collection, we are aware of no way to predict the effect of training-set selection on the effectiveness of the review. The most effective way to validate the adequacy of training-set selection—or any other choice taken in deploying TAR—is a combination of (i) prior scientific validation of the TAR tool, (ii) assurance of its proper application by qualified individuals in a given matter, and (iii) proportionate *post hoc* validation of the end-to-end review process, as discussed in Sections VII and VIII.

There are certainly steps that can be taken to afford the requesting party a measure of insight into and control over the TAR process, and therefore comfort in the judgmental selection of training examples. An obvious way is disclosure by the responding party of the search terms used to find relevant training examples, or the use of search terms specified by (or jointly developed with) the requesting party for this purpose. It is not necessary to include all of the search-term "hits" in the training set; empirical evidence shows that a random sample of the hits is sufficient to "kick-start" the learning algorithm.[56] We believe that "cherry-picking" of training examples is a questionable practice due to its unpredictable impact on the learning algorithm. Therefore, to provide adequate reassurance to the requesting party, all documents examined as potential training examples—however selected—should be coded as responsive or not, and all documents that are coded—both those that are responsive and those that are not—should be included in the training set, not just those that a TAR Whisperer deems to be "good training examples."

Empirical evidence shows that certain active-learning methods, such as CAL, are able to do a thorough job of identifying relevant documents, even when training is accomplished through non-random methods.[57] The judgmental seed set merely provides a hint to get the TAR tool started; it no more "biases" the TAR tool from finding relevant documents than driving in the direction of one's destination "biases" a GPS navigation system from finding the way. While a better seed set may improve the *efficiency* with which active learning discovers relevant documents, our research shows that, even with a seed set selected solely on the basis of a primitive keyword search, active learning is more effective and efficient than the available alternatives.[58]

(footnote continued)

set generated only from an unsophisticated keyword search. *See* Cormack & Grossman, *supra* note 44, at 7, 8.

56. *See id.* at 4.
57. *See generally id.*
58. *See generally id.*

VI. VALIDATION OF TAR OR VALIDATION OF THE END-TO-END REVIEW?

Schieneman and Gricks assert that the TAR process—which they characterize as starting once the collection has been "optimize[d] and manage[d]," and ending once a "review set" has been created[59]—must be validated by ensuring, through statistical sampling, that the review set has achieved certain precision[60] and recall targets whose values are established through a process they refer to as "stabilization."[61]

Not all TAR processes—CAL being a notable exception—involve stabilization, or the creation of a review set. Validation, we argue, should apply to the *end-to-end review*, starting with the original collection and ending with the production set, regardless of which, if any, of the steps are deemed to be "TAR." That is, validation must account for all responsive documents excluded by the review, whether before, during, or after "TAR"; or even when traditional review methods are applied.

Even if each phase, alone, excludes relatively few responsive documents, the combined effect can be substantial. Let us suppose in our hypothetical matter that the enrichment process were to discard 30% of the relevant documents (*i.e.*, were to have a recall of 70%), the TAR process were to have a recall of 75%, and the final human review were to have a recall of 70%. These numbers, in isolation, might be considered reasonable, but consider the combined effect of all three phases on our example. Of 10,000 responsive documents in the original collection, 7,000 (*i.e.*, 70%) would be retained in the enriched collection. Of the 7,000 responsive documents in the enriched collection, 5,250 (*i.e.*, 75%) would be retained in the review set. Of the 5,250 responsive documents in the review set, assuming that none were withheld as privileged, 3,675 (*i.e.*, 70%) would be identified for production. Accordingly, the recall of the end-to-end review effort would be 36.75% (*i.e.*, 3,675 of the original 10,000 responsive documents). This is a far cry from the end-to-end recall results demonstrated at TREC 2009.

We argue that the sequential application of culling methods—in this instance, enrichment, the "technology-assisted review process" (as defined by Schieneman and Gricks[62]), and the subsequent manual review of the review set—will, due to the multiplier effect described above, generally yield inferior recall. Prior enrichment is unnecessary, as discussed in Section IV, and the subsequent human review effort will

---

59. Schieneman & Gricks, *supra* note 1, at 251, 269.

60. We mention precision only in passing because, as noted by Schieneman and Gricks, precision is less informative than recall, and is not likely to be an issue in most reviews. *Id.* at 268 ("As a practical matter, the precision of technology-assisted review processes is much greater than the precision of other review alternatives, making it unlikely that irrelevant documents are being produced for some improper purpose in violation of Rule 26(g). Moreover, the documents identified for production through the technology-assisted review process are generally reviewed before production, further minimizing the production of irrelevant documents in violation of Rule 26(g)." (citations omitted)).

61. *Id.* at 263.

62. *Id.* at 269.

introduce its own error.[63] Whether or not several culling phases are used, validation should estimate how many of the responsive documents in the *original collection* are being *produced*, not how many of the responsive documents in the *enriched collection* are being *reviewed*. Extraordinary efforts to achieve the latter, we argue, often come at the expense of the former.

VII. VALIDATION METHODS

Schieneman and Gricks narrowly equate validation with *post hoc* statistical sampling to estimate recall, excluding other forms of evidence pertinent to the efficacy of a review. Furthermore, they assert, without support, that certain sampling methods provide accurate and efficient recall estimates.

We introduce, in Subsection A, the framework for our argument that prior evidence of a method's effectiveness, coupled with evidence that the method was properly applied by qualified individuals, is at least as important as *post hoc* validation. We argue, in Subsection B, that recall, while an intuitively appealing and reasonably informative measure of review effectiveness, does not tell the whole story, and is difficult to measure accurately or consistently. We show, in Subsection C, that *post hoc* sampling methods for recall estimation in common use today—including those advanced by Schieneman and Gricks, and others—are either mathematically unsound or could require the manual review of unreasonably large samples, resulting in disproportionate effort.

A. Prior and *Post Hoc* Evidence

Validating a review effort involves consideration of all available evidence, including prior scientific evidence confirming the validity of the method, evidence that the method was properly applied by qualified individuals, and subsequent evidence that readily observable phenomena are consistent with a successful outcome.

When cooking a turkey, one can be reasonably certain that it is done, and hence free from salmonella, when it reaches a temperature of at least 165 degrees throughout. One can be reasonably sure it has reached a temperature of at least 165 degrees throughout by cooking it for a specific

---

63. *See generally, e.g.*, Maura R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267 (2012), *available at* http://digitalcommons.pace.edu/plr/vol32/iss2/1/; William Webber, Douglas W. Oard, Falk Scholer & Bruce Hedin, *Assessor Error in Stratified Evaluation*, in PROCEEDINGS OF THE 19TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM '10) 539 (Oct. 2010), *available at* http://www.williamwebber.com/research/papers/wosh10_cikm.pdf; Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70 (2010); Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697 (2000), *available at* http://dx.doi.org/10.1016/S0306-4573(00)00010-8 (hereinafter "Voorhees").

amount of time, depending on the oven temperature, the weight of the turkey, and whether the turkey is initially frozen, refrigerated, or at room temperature. Alternatively, when one believes that the turkey is ready for consumption, one may probe the turkey with a thermometer at various places. Both of these approaches have been validated by biological, medical, and epidemiological evidence. Cooking a turkey requires adherence, by a competent cook, to a recipe that is known to work, while observing that tools like the oven, timer, and thermometer appear to behave properly, and that the appearance, aroma, and texture of the turkey turn out as expected. The totality of the evidence—vetting the method in advance, competently and diligently applying the method, and monitoring observable phenomena following the application of the method—supports the reasonable conclusion that dinner is ready.

B. Recall Uncertainties

Recall can be an informative indicator of the completeness of a review effort, but it is difficult to measure properly and can be misleading. Scientific studies like TREC have expended vast resources—far more than would be reasonable and proportionate in most matters—measuring the recall of various information-retrieval approaches. Even so, the margins of error at the 95% confidence level, used at TREC 2009, and considered in the JOLT Study, ranged from ±5.7% to ±25.9%[64]—larger than the maximum proposed by Schieneman and Gricks. For many matters, it is neither feasible nor necessary to expend the effort to estimate recall with a margin of error of ±5%, at a 95% confidence level—the standard required for scientific publication.

Margins of error and confidence levels quantify only one source of uncertainty—random error, or uncertainty due to chance—in estimating recall and other statistics. A recall estimate, however, is meaningless without an unambiguous characterization of (i) the set of ESI over which the recall is calculated, and (ii) responsiveness. As noted in Section IV, collection enrichment reduces the set of documents subject to review, potentially yielding a higher and easier-to-calculate recall estimate that is not comparable to the recall estimate that would be obtained based on the original collection. At its core, collection enrichment to facilitate validation is itself an exercise in judgmental sampling from which, as Schieneman and Gricks acknowledge,[65] no valid statistical inference can be drawn.

---

64. Hedin et al., *supra* note 21, at 17 Table 6 (Topics 201, 202, 203, and 207). The margins of error may be derived from the confidence intervals ("95% C.I.") shown in Table 6. For a method of computing somewhat smaller margins of error, *see* William Webber, *Approximate Recall Confidence Intervals*, 31 ACM TRANSACTIONS ON INFO. SYS. 1 (Jan. 2013), preprint *available at* http://arxiv.org/pdf/1202.2880v1.pdf (hereinafter "Webber").

65. Schieneman & Gricks, *supra* note 1, at 260 ("A method in which a sample of the document population is drawn, based at least in part on subjective factors, so as to include the 'most interesting' documents by some criterion; the sample resulting from such a method. *Unlike a*

To define any set, it is necessary to define what constitutes an element of the set. In the case of ESI, this choice can be surprisingly complicated and controversial. Are duplicate documents distinct elements of the set, or the same element? Are email messages and their attachments separate elements, or is a message with all of its attachments, one element? Even the TREC 2009 coordinators were unable to resolve the latter issue and reported two sets of substantially different recall estimates:

1. Message-level recall, in which each email message, with attachments, was considered to be a single element. For the efforts reported in the JOLT Study, message-level recall ranged from 67.3% to 86.5%, with an average of 76.8%;[66]

2. Document-level recall, in which each email and its attachments were considered to be distinct elements. For the efforts reported in the JOLT Study, document-level recall ranged from 76.1% to 89.6%, with an average of 84.1%.[67]

Unfortunately, much commentary, many pleadings, and some opinions cite aspirational or measured recall values—70% or 75%—typically citing the TREC 2009 *message-level* results for support, without specifying the set of ESI over which their own recall is calculated (often what appears to be *document-level* recall). This is an invalid comparison.

Defining responsiveness is equally problematic. While requests for production ("RFPs") are typically specified in writing, they are subject to interpretation, and reasonable, informed reviewers may disagree on their interpretation, or on how to apply that interpretation to determine whether or not any particular document is responsive. Even a single reviewer may make a different determination, under different circumstances, including his or her knowledge at the time of the determination, arguments or incentives to decide one way or the other, and human factors such as fatigue or distraction, which can result in clerical errors. Experts disagree surprisingly often with each other, and even with themselves.[68]

Uncertainty in determining responsiveness necessarily limits the ability to measure recall. This limitation can be profound. At TREC 2009, "pre-adjudication scores" (*i.e.*, metrics computed prior to

(footnote continued)

random sample, the statistical properties of a judgmental sample may not be extrapolated to the entire population." (emphasis in original) (quoting *TAR Glossary*, *supra* note 3, at 21)).

66. Hedin et al., *supra* note 21, at 17 Table 6 (Topics 201, 202, 203, and 207).

67. National Institute of Standards and Technology, *Per-Topic Scores: TREC 2009 Legal Track, Interactive Task*, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS, Appendix titled *Legal Interactive Run Results*, at 3, 4, 5, 6, 9 Tables 4, 8, 12, 16, 28 (Document-based, Post-Adjudication Scores for Topics 201, 202, 203, 204, and 207), *available at* http://trec.nist.gov/pubs/trec18/appendices/app09int2.pdf.

68. *See generally supra* note 63.

undertaking the quality-assurance process), including recall, were determined using the coding of third-party reviewers (either law students or contract reviewers). The pre-adjudication message-level recall for the efforts reported in the JOLT Study ranged from 3.6% to 70.9%, with an average of 24.3%[69]—more than 50% lower than the official recall scores reported above after quality-assurance efforts. The official recall scores were achieved through the use of an extraordinary appeal and adjudication process used to correct coding errors by the third-party reviewers. The bottom line is that *inconsistencies in responsiveness determinations limit the ability to estimate recall*. Even a perfect review, performed by an expert—with a recall of 100%—is unlikely to achieve a measured recall of higher than 70%, if the final responsiveness assessment is made by a second, independent expert.[70] If the responsiveness assessment is made by an inexpert reviewer, chances are that the measured recall will be considerably lower, as evidenced by the TREC 2009 "pre-adjudication scores."

There is not necessarily a one-to-one correspondence between the definition of responsiveness and a single written RFP. Typically, a discovery request or subpoena will include at least several, and sometimes dozens, of RFPs. Often, a number of RFPs will seek similar subject matter from the same universe of ESI, and it would therefore be expedient to use a single, combined review (or a small number of combined reviews) to find documents responsive to all of these RFPs. As far as the review is concerned, a document is responsive if it is responsive to at least one of the RFPs. It is not obvious that a single recall estimate based on this amalgamated definition of responsiveness is a good indicator of the thoroughness of the review. Let us suppose that, in our hypothetical matter, the discovery request were to contain three RFPs ("RFP 1," "RFP 2," and "RFP 3"). Suppose further that, of the 10,000 responsive documents, 9,000 were responsive to RFP 1, 900 were responsive to RFP 2, and 100 were responsive to RFP 3. A reported recall of 70% might indicate that 6,300 documents responsive to RFP 1, 630 documents responsive to RFP 2, and 70 documents responsive to RFP 3, were found (*i.e.*, there is 70% recall on each RFP). A reported recall of 70% might equally well indicate that, by accident or design, 7,000 documents responsive to RFP 1, and none responsive to the other two RFPs, were found. In the latter case, "70% recall" presents as reasonable a clearly unreasonable result. We argue that, in order to be

---

69. National Institute of Standards and Technology, *supra* note 67, at 3, 4, 5, 6, 9 Tables 1, 5, 9, 13, 25 (Message-based, Pre-Adjudication Scores for Topics 201, 202, 203, 204, and 207).

70. Maura R. Grossman, Gordon V. Cormack, Bruce Hedin & Douglas W. Oard, *Overview of the TREC 2011 Legal Track*, in NIST SPECIAL PUBLICATION: SP 500-296, THE TWENTIETH TEXT RETRIEVAL CONFERENCE (TREC 2011) PROCEEDINGS, at 9 (2011), *available at* http://trec.nist.gov/pubs/trec20/papers/LEGAL.OVERVIEW.2011.pdf ("Somewhat higher recall may be achieved with more effort, but it is unclear whether improvements in recall measures above 70% are meaningful, given the inherent uncertainties arising from sampling and human assessment of responsiveness."); Voorhees, *supra* note 63, at 701 ("The scores for the two sets of secondary judgments imply a practical upper bound on retrieval system performance of 65% precision at 65% recall since that is the level at which humans agree with one another.").

considered reasonable, the review must be shown to find material responsive to *each* of the RFPs (assuming the collection contains documents responsive to each), which a single recall measure cannot do. On the other hand, separately estimating the recall for each individual RFP is likely to require disproportionate effort. Statistics cannot solve this problem, and therefore other approaches may be warranted.

By similar reasoning, we argue that, to be considered reasonable, a review should be shown to cover all aspects of responsiveness, even if each aspect is not explicitly specified in a separate RFP. For example, if the collection were to include email messages, PDFs, word-processing documents, and spreadsheets containing potentially responsive information, no one would question that the review should cover all of these file types. By the same token, if the RFP were to request documents reflecting "actions by any board member," the review should be performed to cover all board members. In general, a single recall estimate may mask the fact that some identifiable sub-population of documents has been searched inadequately, or missed entirely.

Recall further masks the issue that responsive documents differ in their importance in resolving the issues in dispute. If the information contained in the 30% of the responsive documents omitted by a review achieving 70% recall is largely duplicative or unimportant, the review is qualitatively superior to another with the same recall that omits a large amount of non-duplicative, important information.

Regardless of the recall estimate, it behooves the responding party to investigate whether the documents that are missed form an identifiable sub-population, or are non-duplicative and important. If so, the review should continue. On the other hand, if a review strategy has been shown in advance to be effective, is properly applied by qualified individuals to the matter at hand, and observable indicators are consistent with having adequately searched for all aspects of responsiveness and all identifiable sub-populations within the collection, then a less precise or less burdensome recall estimate that yields a result consistent with a reasonable search should be sufficient. Extraordinary efforts to estimate recall—at the standard required for scientific publication—are unwarranted and disproportionate in light of the totality of the evidence.

C. *Post Hoc* Recall Estimation Methods

Uncertainty in recall estimates arises from both random error (*i.e.*, chance), and also from inconsistency in relevance determinations. An estimate should minimize the overall uncertainty from both sources, for a given level of estimation effort. In the proportionality equation, effort to reduce uncertainty in estimation competes for resources with effort to improve the review process. It is therefore important to use an efficient estimation method, and to expend resources to measure recall only as necessary to satisfy the obligation of a reasonable inquiry under Rule 26(g). Indeed, there may be some situations where it is disproportionate

to compute a statistical recall estimate at all, in light of the cost of sampling and the availability of other methods of validation.

A common way to estimate recall is by *post hoc* analysis, in much the same way that a thermometer is used to determine that a turkey is cooked before it is served. A number of TAR protocols apply statistical analyses repeatedly to track the progress of the review,[71] in much the same way that a thermometer can be used over and over to track the progress of cooking a turkey. Whether employed *post hoc*, or in an ongoing fashion to track progress, many of the estimation methods proposed or commonly in use today are either statistically unsound or, in our opinion, disproportionately burdensome.

Schieneman and Gricks discuss four statistical methods for computing *post hoc* recall estimates:[72] the "Direct Method," proposed by David D. Lewis on behalf of plaintiffs in the *Kleen Products* case,[73] and ordered by consent in the *In re Actos* case,[74] and three methods that estimate recall as the ratio of two separate statistical estimates (together, the "Ratio Methods"). The Direct Method, according to Schieneman and Gricks, may involve disproportionate effort, in light of their assertion that "there are alternative means of calculating recall that do not require such a significant effort"[75] (*i.e.*, the Ratio Methods). While we agree that employing the Direct Method will often entail disproportionate effort, we also argue that the purported advantages of the Ratio Methods are largely illusory, predicated as they are on unsound mathematics.

The Direct Method of estimating recall proceeds as follows. Documents are drawn repeatedly, at random, from the collection, and are reviewed for responsiveness until 385 responsive documents are identified. After—and only after—the 385 responsive documents have been identified, it is determined what percentage of the 385 documents had previously been identified as responsive by the TAR process, however defined. This percentage is the recall estimate, with a margin of error of ±5%, and 95% confidence. The Direct Method is statistically

---

71. Repeated statistical estimates may involve drawing repeated samples (as discussed in Section III), or more commonly, drawing a sample at the outset (a "control set"), and repeating the statistical calculations using the control set as a sample. Problems with the use of repeated estimates to track progress, and to determine when recall is "adequate," are addressed elsewhere. *See*, *e.g.*, Mossaab Bagdouri, William Webber, David D. Lewis & Douglas W. Oard, *Toward Minimizing the Annotation Cost of Certified Text Classification*, in PROCEEDINGS OF THE 22ND ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM '13) 989 (Oct. 2013), *available at* http://dl.acm.org/citation.cfm?doid=2505515.2505708; William Webber, Mossaab Bagdouri, David D. Lewis & Douglas W. Oard, *Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness*, in PROCEEDINGS OF THE 36TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR '13) 933 (July 2013), authors' version *available at* http://terpconnect.umd.edu/~oard/pdf/sigir13webber.pdf.

72. Schieneman & Gricks, *supra* note 1, at 272–73.

73. Hr'g Tr. at 259–64, *Kleen Prods., LLC* v. *Packaging Corp. of Am.*, No. 10-C-5711 (N.D. Ill. Feb. 21, 2002).

74. Case Mgmt. Order: Protocol Relating to the Produc. of ESI, *In Re: Actos (Pioglitazone) Prods. Liab. Litig.*, MDL No. 6:11-md-2299 (W.D. La. July 27, 2012).

75. Schieneman & Gricks, *supra* note 1, at 273.

sound, but is quite burdensome, especially when richness is low. In our hypothetical matter, it would be necessary to draw and manually review about 38,500 documents to calculate the recall estimate—nearly twice as much effort as would be involved in the entire end-to-end CAL process.[76] Were the richness lower—as it was for four of the review tasks considered in the JOLT Study—the burden would be even more onerous. In general, the Direct Method entails the manual review of 385/R documents, where R is the richness of the collection. For each of the two JOLT Study tasks for which richness was 0.3%, the Direct Method would require the review of 385/0.003=128,333 documents to compute the recall estimate. The evidence that would be educed from such an estimate is clearly not worth the candle.

Moreover, if the Direct Method were employed in our hypothetical matter, and were to yield a recall estimate of 75%, it would follow that 25% of the 385 documents—96 in total—would be responsive, but not identified by the TAR process. These documents would obviously need to be produced, but also examined to determine whether they contained non-duplicative, important information. The documents would likely represent a viable seed or training set for a supplemental TAR search. In contrast to sequential culling, the cumulative effect of sequential searching is beneficial. If the supplemental search—applied only to the documents not yet identified as responsive—had a recall of just 60%, it would find 1,500 of the 2,500 remaining responsive documents, for a total of 9,000 documents, or 90% overall recall.

Schieneman and Gricks assert that two Ratio Methods, a basic method (the "Basic Ratio Method") and the method employed in the *Global Aerospace* case[77] (the "*Global* Method"), achieve an acceptable recall estimate, with less sampling effort, than the Direct Method. Although dismissed by Schieneman and Gricks as "unnecessarily indirect,"[78] a third Ratio Method ("eRecall") has been advanced by Herbert L. Roitblat to address the same proportionality concerns.[79] The foundation for these assertions appears to be the following argument:

---

76. *See supra* p. 291 and note 26.

77. Letter from Gordon S. Woodward, Att'y for Landow Entities, to All Counsel, *Global Aerospace Inc.* v. *Landow Aviation, L.P.*, Consol. Case No. CL601040 (Va. Cir. Ct. Nov. 30, 2012) ("'esi' Sampling Report" on docket).

78. Schieneman & Gricks, *supra* note 1, at 272.

79. *Measurement in eDiscovery*, *supra* note 15, at 7–10 ("In order to find 400 responsive documents at 10% Richness or Prevalence, we would have to sample approximately 4,000 documents, randomly chosen from the collection as a whole, without regard to whether they were predicted to be responsive or not (10% of 4,000 is 400). That's a lot of work, and it may be more than the number of documents needed to train the process in the first place (if we are using predictive coding).

If Prevalence is lower, if only a small percentage of documents is actually responsive, measuring Recall directly can be even more costly, because a still large sample of random documents would have to be examined. Calculating Recall directly can be done, but it takes a very substantial amount of work just to find the responsive documents to measure. . . .

Fortunately, there are other ways to assess whether we conducted a reasonable inquiry with far less effort.").

> *The Ratio Method Fallacy*: A sample of size 385 yields a margin of error of ±5%, with 95% confidence. Therefore, if we use a sample of size 385 to estimate the number of responsive documents in the *production*, and another sample of size 385 to estimate the number of responsive documents in the *collection*, the ratio of these two estimates is a valid recall estimate with a margin of error of ±5%, and a confidence level of 95%.

This argument is specious. The Ratio Methods differ from one another only in how they estimate the number of responsive documents in the production, and the number of responsive documents in the collection, prior to dividing the former by the latter to yield an estimate of recall. The Basic Ratio Method samples the *production* and the *collection*;[80] the *Global* Method samples the *production* and the *null set*;[81] while eRecall samples the *null set* and the *collection*.[82]

In each case, the sample estimates are combined using simple algebra to form an estimate of recall. However, in each case, the algebra yields a biased[83] point estimate, with a margin of error dramatically larger than ±5% and/or a confidence level dramatically lower than 95%.

While the correct calculations are complex,[84] it is a straightforward matter to demonstrate that the calculations underlying *The Ratio Method Fallacy* are incorrect. From the definitions of point estimate,[85] margin of error,[86] and confidence level,[87] we know that if the sampling is repeated a large number of times, the average of an unbiased point estimate of recall should approach the true value, and the point estimate should be within ±5% of the true value, 95% of the time. To validate each estimation method, we simulated its application 100,000 times to our hypothetical production set known to have a recall of 75%. If valid, the estimation

---

80. Schieneman and Gricks, *supra* note 1, at 273.

81. *Id.* The "Null Set" is defined as "The set of Documents that are not returned by a search process, or that are identified as Not Relevant by a review process." *TAR Glossary*, *supra* note 3, at 25.

82. *Measurement in eDiscovery*, *supra* note 15, at 7–10. *See also* Herbert L. Roitblat, *A Tutorial on Sampling in Predictive Coding* (OrcaTec LLC 2013), at 3, *available at* http://orcatec.com/2013/10/07/a-tutorial-on-sampling-in-predictive-coding/.

83. A statistic is biased if the long-term average of the statistic is not equal to the parameter it is estimating. A biased estimation method, when repeated a large number of times, will yield an estimate that is, on average, higher or lower than the true value. In contrast, an unbiased estimation method, if repeated a large number of times, will, on average, be indistinguishable from the true value. *See* MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 3 (2d ed. 2001) ("An estimate[e] is said to be unbiased if its average over all possible random samples equals the population parameter no matter what that value may be.").

84. *See generally* Webber, *supra* note 64.

85. "Point Estimate" is defined as "The most likely value for a Population characteristic. When combined with a Margin of Error (or Confidence Interval) and a Confidence Level, it reflects a Statistical Estimate." *TAR Glossary*, *supra* note 3, at 25.

86. "Margin of Error" is defined as "The maximum amount by which a Point Estimate might likely deviate from the true value, typically expressed as 'plus or minus' a percentage, with a particular Confidence Level." *Id*. at 22.

87. "Confidence Level" is defined as "[T]he chance that a Confidence Interval derived from a Random Sample will include the true value." *Id.* at 12.

method should yield point estimates of recall whose average is nearly indistinguishable from 75%,[88] of which 95% fall between 70% and 80% (*i.e.*, within the margin of error of ±5%). Moreover, if the point estimate were unbiased, and the margin of error were ±5% as claimed, the 2.5th percentile of the estimates should be 70%, while the 97.5th percentile should be 80%. In other words, 2.5% of the time, the estimate should fall below 70%, and 2.5% of the time it should fall above 80%.

As seen in Tables 2 and 3, the Direct Method yields an unbiased estimate, with more than 95% of the estimates falling within the predicted margin of error. The 2.5th and 97.5th percentiles are close to the predicted values. In short, the Direct Method is statistically sound, yielding an unbiased point estimate and a margin of error slightly better than claimed.

The Ratio Methods, on the other hand, yield biased estimates with margins of error vastly larger than claimed. The Basic Ratio Method and the *Global* Method overestimate—while eRecall underestimates—the true recall value, especially, but not exclusively, when prevalence is low. None of the Ratio Methods provides recall estimates that fall within the required margin of error anywhere near 95% of the time. None of the Ratio Methods has 2.5th and 97.5th percentiles anywhere near 70% and 80%, respectively. In short, the appeal of less sampling effort[89] is belied by the fact that the estimates are unsound. On the other hand, the Direct Method, while mathematically sound, entails *much* larger samples, especially for low prevalence. In short, there is no free lunch.

Table 2: Recall Estimation Validity Results—1% Collection Richness

Method	Sample Size Required for Computation of Recall	Average (Should Be 75%)	% Within ±5% (Should Be 95%)	2.5th Percentile (Should Be 70%)	97.5th Percentile (Should Be 80%)
Direct	38,500	75.0%	97.5%	70.6%	79.2%
Basic Ratio	770	83.4%	17.4%	35.7%	100.0%
Global	770	79.4%	36.7%	48.5%	100.0%
eRecall	770	69.3%	8.9%	0.0%	100.0%
Simulated recall estimates for a hypothetical review effort with a known recall of 75%, a known precision of 83.3%, and a known prevalence of 1%.					

88. STEVEN K. THOMPSON, SAMPLING, 34 (3d ed. 2012) ("One can assess the bias or expected error by looking at the difference between the average value of the estimator over all the samples and the true test population characteristic.").

89. Schieneman & Gricks, *supra* note 1, at 273.

Table 3: Recall Estimation Validity Results—10% Collection Richness

Method	Sample Size Required for Computation of Recall	Average (Should Be 75%)	% Within ±5% (Should Be 95%)	2.5th Percentile (Should Be 70%)	97.5th Percentile (Should Be 80%)
Direct	3,850	75.0%	97.5%	70.6%	79.2%
Basic Ratio	770	76.5%	33.7%	56.8%	100.0%
Global	770	75.4%	61.5%	64.6%	86.6%
eRecall	770	74.4%	43.4%	54.7%	88.9%
Simulated recall estimates for a hypothetical review effort with a known recall of 75%, a known precision of 83.3%, and a known prevalence of 10%.					

The above examples serve to illustrate that there is no shortcut to achieving a scientific-publication-quality estimate of recall, a standard that may not be necessary or appropriate for the typical TAR review. The bottom line is that better recall estimates require more work; work that may be disproportionate to the contribution of the estimate for validation purposes. The Ratio Methods—the *Global* Method, in particular—can be adjusted to calculate an unbiased point estimate, with an appropriate confidence interval,[90] at the expense of requiring sample sizes comparable to those of the Direct Method.

It bears repeating that, regardless of sample size or sampling regimen, inconsistency in human assessment of responsiveness can result in substantial error in recall estimates, as illustrated in Subsection B. Sample-based estimation depends on human review of the sample, and even an expert reviewer will be fallible, resulting in estimation error.[91] With perhaps disproportionate effort, this source of estimation error can be reduced, but not eliminated; for example, by using a committee of experts to review the sample. With additional effort, statistical estimation error may likewise be reduced by increasing sample size indefinitely. The bottom line is that efforts to reduce one kind of error come at the expense of efforts to reduce the other, and, in any event, validation efforts compete for resources that might otherwise be used to conduct a better review.

VIII. A WAY FORWARD: COMBINING PRIOR AND *POST HOC* EVALUATION

We are unconvinced that extraordinary efforts to conduct *post hoc* recall estimates are justified for the purpose of validating a single review, for the same reason that it is not necessary to use a laboratory-quality

90. *See generally* Webber, *supra* note 64.
91. *See supra* note 70. *See also generally supra* note 63.

thermometer—or indeed any thermometer at all—to ensure with reasonable certainty that a turkey is cooked. We believe that effort spent in advance to scientifically validate the TAR method, ongoing oversight to ensure the method is properly applied by qualified individuals, and proportionate *post hoc* testing, would provide more confidence in the review effort than Herculean *post hoc* sampling efforts on a per-matter basis. In so arguing, we consider the factors summarized in the syllabus of the *Daubert* case:[92]

> Faced with a proffer of expert scientific testimony under Rule 702, the trial judge, pursuant to Rule 104(a), must make a preliminary assessment of whether the testimony's underlying reasoning or methodology is scientifically valid and properly can be applied to the facts at issue. Many considerations will bear on the inquiry, including whether the theory or technique in question can be (and has been) tested, whether it has been subjected to peer review and publication, its known or potential error rate, and the existence and maintenance of standards controlling its operation, and whether it has attracted widespread acceptance within a relevant scientific community. The inquiry is a flexible one, and its focus must be solely on principles and methodology, not on the conclusions that they generate.

To be clear, we do not suggest that a formal *Daubert* hearing is necessary or appropriate in evaluating individual TAR efforts; the admissibility of expert testimony at trial and the adequacy of production present different objectives and standards.[93] Nevertheless, both share the common concern of assessing whether a method of inquiry is reasonable, valid, and properly applied. Accordingly, even if not dispositive, the *Daubert* factors may be instructive in the TAR context. Notably, the *Daubert* test focuses on *a priori* validation—establishing, through accepted scientific standards, the applicability of the method to the task at hand, its potential error rate, and standards controlling its operation. Therefore, TAR tools and methods should be pre-established as valid in their own right, not re-established *de novo* for each and every matter. *Daubert* further emphasizes flexibility, which we interpret to mean that *all available evidence* should be considered in validating a TAR method

---

92. *Daubert* v. *Merrell Dow Pharm., Inc.*, 509 U.S. 579, 580 (1993).

93. *See Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, 189 (S.D.N.Y. 2012) ("[Federal Rule of Evidence] 702 and *Daubert* simply are not applicable to how documents are searched for and found in discovery."); *but see* David J. Waxse & Brenda Yoakum-Kriz, *Experts on Computer-Assisted Review: Why Federal Rule of Evidence 702 Should Apply to Their Use*, 52 Washburn L.J. 207, 223 (Spring 2013), *available at* http://contentdm.washburnlaw.edu/cdm/ref/collection/wlj/id/6195 ("Rule 702 and the *Daubert* standard should be applied to experts with technical expertise or knowledge pertinent to a party's ESI search and review methodologies and who provide the court with evidence on discovery disputes involving these methods.").

and its application to any particular review—not just a single bright-line test such as recall.

Notwithstanding the caveats expressed above with regard to recall—or any statistical measure—as the sole gauge of success, recall can be estimated prior to any particular review effort, based on the results of previous efforts. For example, the CAL method examined in the JOLT Study achieved document-level recall point estimates of 84.3%, 84.4%, 86.0%, and 89.6% on the four TREC 2009 review tasks to which it was applied.[94] Applying the t-distribution[95] to these estimates yields a 95% confidence interval of 82.1% to 90.0%, meaning that, if the same method were applied with equal diligence to a fifth review task, we would expect with 95% confidence to achieve, on average, a true recall between 82.1% and 90.0%. In other words, *a priori* statistical evaluation of the method on four tasks can yield as high a confidence level in estimating the result of a fifth task, as can *post hoc* evaluation of the fifth task alone. In addition, a recall point estimate computed at the end of a review—even one with a high margin of error—augments confidence, not only in the particular result, but in future results. Once a sufficient *a priori* confidence level has been established, it should be sufficient to ensure that the method is properly applied by qualified individuals, and that readily observable evidence—both statistical and non-statistical—is consistent with the proper functioning of the method.

We do not claim that the use of CAL, prior scientific validation of the TAR method, competence and oversight in its application, and proportionate *post hoc* sampling are the only ways forward, but we are unaware of any approaches that work better, either for TAR or for validation. There is much more to be learned about the application and validation of TAR, and, we argue, the legal industry would be better served by researching, developing, and implementing improvements in TAR, than by expending heroic efforts—under the guise of compliance with Rule 26(g)—to validate individual TAR efforts through the application of statistical methods that are either mathematically unsound, or so disproportionate that they create disincentives to use TAR, and threaten to eradicate the savings afforded by its use.

IX. CONCLUSION

Any effort to codify what is reasonable—under Rule 26(g) or otherwise—is necessarily limited by the difficulty of anticipating the unintended consequences of such prescriptions. We have illustrated a number of circumstances in which an obligation to follow the practices dictated by Schieneman and Gricks would preclude the use of perfectly reasonable—indeed superior—TAR methods and validation strategies.

---

94. *See supra* note 67. *See also* Cormack & Mojdeh, *supra* note 10, at 8 Table 5.

95. STEFAN BÜTTCHER, CHARLES L. A. CLARKE & GORDON V. CORMACK, INFORMATION RETRIEVAL: IMPLEMENTING AND EVALUATING SEARCH ENGINES 423 (2010).

An obligation to follow Schieneman and Gricks' prescriptions would entrench one particular approach to TAR, and stifle the state of the art.

There is no reason that the criteria for reasonableness under Rule 26(g) should be "unique" to TAR: All review methods leave responsive documents behind, and all review methods may be subject to validation—statistical or otherwise. Application of standards of reasonableness piecemeal to discrete phases of a review effort—*e.g.*, collection, disclosure, training, stabilization, and validation—has the potential to invite discovery about discovery and successive motions practice, thereby impeding the "just, speedy, and inexpensive determination of every action and proceeding."[96]

Accordingly, we believe that the Rule 26(g) requirement to conduct a reasonable inquiry may be fulfilled without undertaking the burdensome and potentially counterproductive enrichment, training, stabilization, and validation techniques asserted by Schieneman and Gricks to be obligations under Rule 26(g). Exceptional efforts to enrich the corpus have not been shown to enhance either review quality or confidence in the result. Random selection of training documents, to the exclusion of non-random methods, has not been shown to enhance either review quality or confidence in the result; to the contrary, Continuous Active Learning—a method independently shown to achieve superior results—uses a judgmental seed set and active learning, both non-random methods, and is entirely consistent with the obligations mandated by Rule 26(g). Finally, heroic efforts to measure recall for any particular review may be unduly burdensome and divert resources from conducting a better review. Such metrics can be gamed or misinterpreted, and will often require statistical sophistication beyond what might reasonably be expected of the average responding party. Most dangerous of all, however, is the risk that efforts aimed at maximizing recall estimates may drive TAR workflow choices in a direction that is inconsistent with best practice.

96. Fed. R. Civ. P. 1.

EXHIBIT C

This article has been accepted for publication in Volume 18, Issue 1 of the Ohio State Technology Law Journal (December 2021)

The eDiscovery Medicine Show

Maura R. Grossman, J.D., Ph.D.* and Gordon V. Cormack, Ph.D.*

As recently as 100 years ago, harmful practices such as bloodletting were still advanced as sound medical practice by expert practitioners.[1] Bloodletting gradually fell into disfavor as a growing body of scientific evidence showed its ineffectiveness and demonstrated the effectiveness of various pharmaceuticals for the prevention and treatment of certain diseases. Basking in the reflected glory of such scientifically proven medicines, unscrupulous purveyors of magical elixirs promoted their wares using pseudo-scientific evidence and testimonials from quacks and charlatans, presented along with free entertainment. These medicine shows persisted until, among other things, the Food and Drug Administration was given the authority to prosecute unsubstantiated therapeutic claims in 1938.[2]

eDiscovery methods, like therapeutics, are amenable to scientific evaluation. But practitioners and their "experts," vendors, and clients often ignore empirical evidence, citing instead existing or past practice to justify, for example, culling electronically stored information ("ESI") using untested search terms, establishing neither their necessity nor their efficacy. Or, they use pseudo-science to promote various potions marketed as "Artificial Intelligence," "AI," "technology-assisted review," or "TAR." Or, they employ pseudo-science and various logical fallacies to impugn scientific studies that contradict their claims. Or they point to the oft-cited Sedona Principle 6[3] as justification to do whatever they please. Or, sometimes, even all of the above. Trade shows and other "educational" activities sponsored by vendors

---

* Maura R. Grossman, J.D., Ph.D., is a Research Professor, and Gordon V. Cormack, Ph.D., is a Professor, in the David R. Cheriton School of Computer Science at the University of Waterloo, in Ontario, Canada. Professor Grossman is also Principal of Maura Grossman Law, an eDiscovery law and consulting firm in Buffalo, New York, U.S.A. Professor Grossman's work is supported, in part, by the National Sciences and Engineering Council of Canada ("NSERC"). The opinions expressed in this piece are the authors' own and do not necessarily reflect the views of the institutions, organizations, or clients with which they are affiliated. The authors wish to thank Jason R. Baron for his thoughtful comments on earlier drafts of this article.

[1] *See* William Osler, *The principles and practice of medicine: designed for the use of practitioners and students of medicine* (D. Appleton, 1892). *But see* Charles S. Bryan, *New observations support William Osler's rationale for systemic bloodletting*, in 32 Baylor U. Med. Ctr. Proc. 372-76 (Taylor & Francis, 2019), https://www.tandfonline.com/doi/full/10.1080/08998280.2019.1615331 (arguing that Osler's prescribed indications might be considered rational within the context of 19th century medicine).

[2] *See* Federal Food, Drug, and Cosmetic Act (June 24, 1938).

[3] Sedona Principle 6 states that "[r]esponding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information." The Sedona Conference, *The Sedona Conference Principles, Third Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production*, 19 Sedona Conf. J. 1, 118-30 (2018), https://thesedonaconference.org/sites/default/files/publications/The%20Sedona%20Principles%20Third%20Edition.19TSCJ1.pdf.

promote their wares, complete with pseudo-scientific results, testimonials, sponsored receptions, prizes, and hospitality suites. The Continuing Legal Education ("CLE") industry and the trade press often echo these testimonials, failing to discriminate between practice and sound practice—let alone best practice—or between science and pseudo-science. So far, neither the courts nor any other authority has taken up the mantle, leaving parties to fend for themselves in the eDiscovery Wild West.

The eDiscovery task is one of information retrieval, which has been a discipline of scientific study for more than 70 years.[4] Those 70 years have seen considerable advances in information retrieval methods, as well as ways to measure their effectiveness. One of the earliest and most primitive approaches to information retrieval is the so-called Boolean search, which retrieves only documents containing certain search terms or combinations of search terms, as specified by a manually constructed query. "Keyword culling" in eDiscovery almost always employs Boolean search, with minor enhancements that take into account the order and proximity of the terms within a document. Parties sometimes exchange "hit reports," quantifying the number of times the search terms appear in the document collection, but these reports provide little, if any, useful information about the quality of the search terms.

The most common—but certainly not the only—way to quantify the effectiveness of an information retrieval effort is to estimate recall and precision. These measures rely on the convenient fiction of binary relevance: that there is a "ground truth," and that any particular document either is or is not relevant to the information need that occasioned the information retrieval effort. If we could somehow know with certainty the relevance of every document in the collection to be searched, recall would be the percentage of all relevant documents that were retrieved, while precision would be the percentage of all retrieved documents that were relevant.[5] But we cannot know with certainty—or even with high confidence—the relevance of every document, or even the number of relevant documents in the collection, and thus recall. At best, we can estimate them with varying—and typically large—margins of error. Any method whose operation depends on a purportedly precise estimate of recall involves pseudo-science and bogus statistics.

While some leading information retrieval scientists eschew the use of recall altogether,[6] it is generally accepted that recall and precision can be used to gauge the *relative effectiveness* of competing information retrieval methods, provided that recall is measured under the same conditions, using the same information need, the same document collection, and the same independently derived relevance assessments.[7] The National Institute of Standards and Technology's ("NIST's") Text REtrieval Conference ("TREC") was founded in 1992, precisely to mount a heroic collaborative effort, among academic, industry, and government researchers, to create such conditions—conditions that cannot possibly be

[4] *See* Mark Sanderson and W. Bruce Croft. *The history of information retrieval research*, 100 Proc. of the IEEE 100 Special Centennial Issue 1444-51 (2012), https://ieeexplore.ieee.org/iel5/5/4357935/06182576.pdf.
[5] *See* Maura R. Grossman and Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 Fed. Cts. L. Rev. 1, 25, 27 (2013), https://www.fclr.org/fclr/articles/html/2010/grossman.pdf.
[6] *See* Justin Zobel, Alistair Moffat, and Laurence A.F. Park, *Against recall: is it persistence, cardinality, density, coverage, or totality?*, in 43 ACM SIGIR Forum 3-8 (2009), https://dl.acm.org/doi/10.1145/1670598.1670600.
[7] *See* Ellen M. Voorhees, *Variations in relevance judgments and the measurement of retrieval effectiveness*, 36 Info. Processing & Mgmt. 697-716 (2000),), https://doi.org/10.1016/S0306-4573(00)00010-8.

reproduced within the context of a particular eDiscovery effort.[8] As such, efforts like TREC are useful to compare the effectiveness and reliability of different information retrieval methods, not to determine the effectiveness of any particular information retrieval effort. Under TREC-like conditions, 65% recall and 65% precision represents a good result—approximately what we would expect if the entire collection were assessed by one subject matter expert ("SME"), with recall and precision estimated using the independent assessments of a second SME.[9]

65% recall and 65% precision are not necessarily the best possible results. If, for example, the majority vote of independent assessments by three SMEs were used, they might achieve on the order of 75% recall and 75% precision, according to the same evaluation conditions.[10] For similar reasons, it is possible for certain specific TAR methods—where relevance is determined using input from both an SME *and* artificial intelligence—to achieve better effectiveness than a single SME alone.[11]

Much ink has been spilled over the entirely irrelevant question of whether or not Boolean search is a form of TAR. Boolean search, like TAR, is an information retrieval method. Boolean search using manually constructed queries (as, for example, employed for keyword culling in eDiscovery), unlike certain TAR methods, generally cannot achieve 70% recall and 70% precision.[12] Perhaps one Boolean query can yield 70% recall and 20% precision, while another for the same information need can yield 20% recall and 70% precision. But finding a single query that can yield both 70% recall and 70% precision is likely impossible.

To make matters worse, eDiscovery medicine shows frequently promote the sequential use of two or more information retrieval methods, including Boolean search, TAR, and manual review. The net effect of this concoction is to achieve considerably lower recall than any of its constituent parts: When multiple information retrieval methods are used in sequence, overall recall is the product of the recall for each constituent method. If keyword culling were to achieve 70% recall, the TAR tool were to

---

[8] *See* Ellen M. Voorhees and Donna K. Harman (eds.), 63 *TREC: Experiment and evaluation in information retrieval* (MIT Press, 2005). *See also* TREC Proc., 1992–present, *available at* https://trec.nist.gov/proceedings/proceedings.html (last visited Sept. 28, 2021).

[9] *See* Ellen M. Voorhees, *supra* n.7.

[10] The majority vote of three independent assessments is more likely to agree with another independent assessment than any of the individual assessments. Thus, when another independent assessment is used to evaluate recall and precision, both of these measures will likely be higher for the majority vote than for any individual. This is an application of the well-known statistical phenomenon referred to as *regression to the mean*, "first noted by Sir Francis Galton that 'each peculiarity in man is shared by his kinsmen, but on the average to a less degree.'" Brian S. Everett and Anders Skrondal, *The Cambridge Dictionary of Statistics, Fourth Edition* (Cambridge Univ. Press 2010), at 363, http://www.stewartschultz.com/statistics/books/Cambridge%20Dictionary%20Statistics%204th.pdf.

[11] *See* Gordon V. Cormack and Maura R. Grossman, *Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me*, in Proc. of the 40th Int'l ACM SIGIR Conference on Research and Dev. in Info. Retrieval, 5-14 (2017), https://dl.acm.org/doi/10.1145/3077136.3080812; Maura R. Grossman and Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 Rich. J. L. & Tech. 1 (2011), https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=1344&context=jolt.

[12] *See* Eero Sormunen, *Extensions to the STAIRS study—empirical evidence for the hypothesised ineffectiveness of Boolean queries in large full-text databases*, 4 Info. Retrieval 257-73 (2001), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.1144.

achieve 80% recall, and manual review were to achieve 75% recall, the recall of a review effort combining them in sequence would be 70%×80%×75%=42%. It is possible to quibble with the numbers presented here, but not with the fact that each constituent part is imperfect, and that overall or end-to-end recall is considerably less than the weakest link in the chain.[13]

When applied sequentially, information retrieval methods—whether Boolean search, TAR, or manual review—will always yield inferior recall. Yet the medicine shows would have us believe that we need to consider only the TAR-tool component in our recall calculations, ignoring relevant documents excluded by keyword culling and/or by post-TAR manual review. This is, at best, an extreme case of moving the goalposts, but more likely, a form of legerdemain.

No doubt purveyors and pundits will attack these arguments and examples with special pleading, *argumentum ad hominem*, appeals to common practice or Sedona Principle 6, strawmen, testimonials, and more pseudo-science—anything but a valid scientific experiment published in a peer-reviewed venue.

The following sections outline several of the methods of deception that have been used—and will continue to be used—to promote inferior methods and to cast aspersions on superior ones, until someone is ready to bell the cat.

Misapplication of Effectiveness Measures

Statistical measures are commonly used to evaluate the efficacy of medical treatments as well as information retrieval methods. A cancer treatment might be judged by its five-year survival rate, while an antibiotic might be judged by the probability that it will cure a specific infection. Neither of these statistics can possibly be measured on a case-by-case basis at the time of treatment; they are instead used to establish which treatments have the best chance of success, so that a reasonable choice can be made when the need arises, considering not only efficacy but cost, availability, side effects, and so on. On a case-by-case basis, observations like symptom abatement, temperature reduction, and the absence of a rash are indicators—but not proof—of successful treatment. These indicators of success often occur within a day or two of antibiotic treatment, but it generally takes several more days to eliminate the infection. If treatment is discontinued prematurely, the infection will likely return, in a form more resistant to the treatment than before. It would cause genuine harm to continue antibiotics only so long as symptoms persist, and then to repeat the treatment when symptoms recur. It might be possible to conduct a laboratory blood test to see with reasonable certainty when the infection was gone, but the test itself would entail more burden, cost, and delay than simply completing the prescribed course of antibiotics.

Recall and precision were conceived by scientists to measure the *comparative effectiveness* of information retrieval methods, not the success of a particular information retrieval effort, and especially not as the sole determinant of when a particular information retrieval effort may be discontinued. It is infeasible to compute a precise and accurate estimate of recall within the context of a particular

[13] *See* Maura R. Grossman and Gordon V. Cormack. *Comments on "The implications of Rule 26 (g) on the use of technology-assisted review,"* Fed. Cts. L. Rev. 285, 293-95 (2014), https://www.fclr.org/fclr/articles/pdf/comments-implications-rule26g-tar-62314.pdf.

eDiscovery effort,[14] and even if such an estimate were feasible, it would not inform the question of whether or not the review, if continued, would find enough additional relevant documents to justify the additional effort, a proportionality consideration that is included in the description of the scope of discovery set forth in Federal Rule of Civil Procedure 26(b)(1), and in the "reasonable inquiry" requirement of Federal Rule 26(g)(1).

The conflation of efficacy and success has led to the absurd propositions that no treatment or information retrieval method can be determined to be reasonable or unreasonable in advance, that no irreparable harm occurs when a treatment or search method fails in a particular case, and that success in a particular case can be solely determined by an estimate of a summary measure like body temperature or recall, without regard to symptoms or the quality of the production.

Measuring the Wrong Quantity (a/k/a Searching Under the Streetlight)

Recall and precision measure the *end-to-end effectiveness* of an information retrieval effort. As noted above, the scientific literature indicates that 70% to 80% recall is achievable by certain TAR methods. Many TAR methods employ a software tool in conjunction with manual review and/or prior Boolean keyword culling. The recall of the software tool, in isolation, must be considerably higher—perhaps 85% to 95%—to achieve an overall recall of 70% to 80% for the end-to-end search and review process, when correctly measured with respect to a blind independent assessment.

Non-Blind Assessment

Several of the most egregious applications of pseudo-science involve the use of non-blind experiments. Clever Hans was a horse purported to be able to perform intellectual tasks including arithmetic.[15] In fact, Hans' handler was either deliberately or inadvertently telegraphing the correct answer to Hans through non-verbal cues. A panel of 13 experts debunked the handler's claims by conducting a blinded experiment in which Hans was isolated from the questioner and spectators—Hans literally wore blinders—using questions for which the questioner did and did not know the answers in advance. The experiments concluded that Clever Hans could intuit the correct answer from the questioner, even if the questioner was not Hans' handler, but only if the questioner knew the answer in advance. So Clever Hans was actually less clever than initially thought.

Notwithstanding this debunking, the Clever Hans show continued to tour Germany, attracting large and enthusiastic crowds. Nowadays, recall is calculated from the non-blind relevance assessments of armies of Clever Hanses, to the delight of large and enthusiastic crowds at eDiscovery medicine shows.

Misleading Statistics

Statistics play several roles in information retrieval. First and foremost, statistics are used to calculate the efficacy and reliability of information retrieval methods over many information needs and/or

[14] *See* David C. Blair, *STAIRS redux: Thoughts on the STAIRS evaluation, ten years after*, 47:1 J. Am. Soc. for Info. Sci. 4-22 (1996), https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199601%2947%3A1%3C4%3A%3AAID-ASI2%3E3.0.CO%3B2-3.

[15] *See* Oskar Pfungst, *Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology*," Translation by Carl. L Rahn (2010), *available at* https://www.gutenberg.org/files/33936/33936-h/33936-h.htm#CHAPTER_IV (last visited Sept. 28, 2021).

document collections. Second, summary measures like recall and precision can be estimated for a particular information retrieval effort from relevance assessments for only a statistical sample of the collection. Third, many information retrieval methods rely on statistical algorithms to score and rank documents by their likelihood of relevance. While only the second use has any operational impact on eDiscovery, it is frequently conflated with the other two, and all are conflated with technology-assisted review.

Measuring recall and precision—whether to establish the efficacy and reliability of a method or to assess the quality of a particular review—has little to do with the technology employed for the review, be it Boolean keyword culling and/or TAR and/or manual review. Regardless of what technologies are employed, recall is the fraction of relevant documents *in the entire collection* that are retrieved by the overall effort. On the other hand, the mechanics of statistical information retrieval algorithms are of no more concern to the eDiscovery practitioner than the biochemistry of a pharmaceutical is to the patient.

Part of the reason that these uses are conflated is that TAR was the first eDiscovery technique whose efficacy was evaluated and compared to that of manual review using the tools of information retrieval. TAR practitioners exposed to these methods naturally assumed that the use of statistics was unique to TAR, and that statistical expertise was needed to employ TAR. Purveyors and pundits have harnessed this misconception to suggest that only the TAR tool should be subject to validation, while keyword culling and manual review should be exempt, as they have always been.

At the same time, purveyors and pundits have grossly exaggerated the accuracy with which recall and precision (even of the TAR tool alone) can be estimated using sampling.[16] One tell that is often seen in vendor and pundit publications is the use of the non-sequitur "statistically significant sample." A more subtle mangling of statistical concepts is the notion of a sample having a given margin of error at a given confidence level; the most common examples are "margin of error of ±5% at a 95% confidence level," and "margin of error of ±2% at a 95% confidence level," when a more direct statement of what is really meant is "a random sample of size 385," and "a random sample of size 2,400," respectively. Specifying sample size obliquely by the margin of error that can purportedly be derived from it serves to obfuscate the sampling process and to intimidate practitioners, while strongly implying that the estimates derived from such samples are more precise than they are, and that they are sufficient to yield estimates of *recall* with the stated margins of error. They are not.

At best, samples of size 385—or even 2,400—can give a coarse estimate of recall with a margin of error several times larger than implied, which may identify a major failure, but cannot guarantee success. Furthermore, if the samples are assessed by Clever Hanses who know or can intuit how the documents were previously coded, we can place no bound on the error—and therefore the over- or underestimation—that can result.

The Fallacies of Incredulity, Special Pleading, and Assuming the Converse

Pundits' and practitioners' incredulity that large numbers of responsive documents are excluded by keyword culling and by reviewer assessment errors does not refute the extensive body of scientific evidence showing that they are. Nor does a special pleading argument that their own unique assessors

[16] *See* Maura R. Grossman and Gordon V. Cormack *supra* n.13, at 305-10.

or quality assurance process or bloodletting instruments would have achieved a different result. Nor does pointing out limitations—whether perceived or real—in scientific studies constitute evidence for the converse of the studies' conclusions. There is a strong body of scientific evidence showing that keyword culling removes a substantial number of responsive documents, that certain TAR methods achieve as good or better recall and precision than human assessors, and that non-blind assessments provide unreliable results. If any of the issues raised in arguments to the contrary had merit, they could surely be demonstrated through scientific studies whose results were subject to peer review. That has not happened.

The Ball is in The Courts' Court

Just as bloodletting came to be considered unreasonable in the face of mounting scientific evidence, so too should certain common eDiscovery practices. Just as the snake oils of the early 20th century were unreasonable from the outset, so too are various TAR tools and methods whose efficacy has not been established.

What is needed is a full-scale evaluation, and judicial recognition of which methods are reasonable and which are not. Unfortunately, an evidentiary hearing sufficient to establish reasonableness would likely entail burden and cost disproportionate to the needs of any particular case, as evidenced by the *Kleen Products* case,[17] in which then-Magistrate (now retired) Judge Nan R. Nolan, after two full days of evidentiary hearings and 11 status hearings and Rule 16 conferences[18] with the parties, ordered them essentially to work it out for themselves.

In the near-decade since *Kleen Products*, the scientific body of evidence regarding how to determine the effectiveness and reliability of TAR methods has grown substantially. It is time to revisit what is reasonable in light of this evidence.

Federal Rule of Civil Procedure 26(g)(1) requires that the producing party or their attorney personally certify that "after reasonable inquiry," to the best of their knowledge, information, and belief, the production is not unreasonable, considering the burden or expense, the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action.

Notwithstanding Sedona Principle 6's prescription that "[r]esponding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information," Federal Rule of Civil Procedure 26(g)(1)—which is the law—mandates that this evaluation must be consistent with a "reasonable inquiry," and that the chosen procedures, methodologies, and technologies must be calculated to yield a reasonable production.

The courts need not presume that any method used in the past or any method chosen by the producing party is reasonable, especially if presented with clear and convincing evidence that it is unreliable or ineffective, compared to readily available alternatives. In the face of such evidence, the court need not wait until the method's unreliability or ineffectiveness is manifested in the matter before it, which may occur only at the eleventh hour when proper remediation is no longer practical. At the same time,

[17] *See Kleen Prods. LLC* v. *Packaging Corp. of Am.*, Civ. Ac. No. 10-C-5711, 2012 WL 4498465, at *4 (N.D. Ill. Sept. 28, 2012).
[18] *See* Fed. R. Civ. P. 16(a).

courts need not impose on producing parties the prescriptions of requesting parties unless they are similarly supported by clear and convincing evidence.

In assessing whether an inquiry, an eDiscovery method, or a production is reasonable, the courts may consider the *Daubert* factors:

> Many considerations will bear on the inquiry, including whether the theory or technique in question *can be (and has been) tested*, whether it *has been subjected to peer review and publication*, its *known or potential error rate*, and the *existence and maintenance of standards controlling its operation*, and whether it has *attracted widespread acceptance within a relevant scientific community*. (emphasis added).[19]

To be clear, we do not advocate a full-blown evidentiary hearing every time a TAR or validation process is challenged. Rather, we suggest that the *Daubert* factors offer useful guidance in determining what a reasonable process is pursuant to Fed. R. Civ. P 26(g)(1), and what is proper evidence to this effect.

The scientific literature contains a growing body of peer-reviewed and widely accepted evidence of the effectiveness and error rates of the information retrieval methods used in eDiscovery, including the error rates for Boolean search, manual review, specific TAR tools and protocols, and specific validation methods. eDiscovery and/or validation efforts can be considered tested only when the tools, methods, and standards of operation they employ are comparable to those of the research studies from which supporting evidence is derived.

Merely labeling an eDiscovery method as "TAR" does not establish that it has been tested. Nor does using a TAR tool that has been tested establish that it has been used according to a tested protocol. Nor does naming a particular protocol such as "Continuous Active Learning" or "CAL" establish that the protocol has been followed according to established standards of operation. Nor does proclaiming "75% recall" establish that recall has been calculated according to scientifically established methods with a known error rate, or that such a computation is evidence of either a reasonable inquiry or a reasonable production.

In contrast, producing parties should show—and the courts should demand that they show—the reasonableness of their eDiscovery search and review processes, as well as the resulting production, by hewing closely to tools, methods, and procedures that have been scientifically vetted and shown to be valid and reliable. Anything else belongs in a medicine show.

---

[19] *See* Syllabus (c) to *Daubert* v. *Merrell Dow Pharm., Inc.*, 509 U.S. 579, 580 (1993). The *Daubert* factors are incorporated in Federal Rule of Evidence 702 governing expert testimony. While it is unlikely that Federal Rule of Evidence 702 per se governs eDiscovery representations, there can be no question that the effectiveness and reasonableness of eDiscovery methods is a matter of scientific inquiry. Indeed, in *Victor Stanley, Inc.* v. *Creative Pipe, Inc.*, 250 F.R.D. 251, 260 n.10 (D. Md. 2008), then-Magistrate (now District) Judge Grimm noted that "[i]t cannot credibly be denied that resolving contested issues of whether a particular search and information retrieval method was appropriate—in the context of a motion to compel or motion for protective order—involves scientific, technical or specialized information. If so, then the trial judge must decide a method's appropriateness with the benefit of information from some reliable source—whether an affidavit from a qualified expert, a learned treatise, or, if appropriate, from information judicially noticed. To suggest otherwise is to condemn the trial court to making difficult decisions on inadequate information, which cannot be an outcome that anyone would advocate." We agree.